Search for the production of a Higgs boson decaying into a pair of bottom quarks in association with a pair of top quarks at 13 TeV with the ATLAS detector

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium
(Dr. rer. nat.)

im Fach Physik

eingereicht an der

Mathematisch-Naturwissenschaftlichen Fakultät

der Humboldt-Universität zu Berlin

von

Ing. Filip Nechanský

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter:

1. PD Dr. Klaus Mönig Phd.
2. Prof. Dr. Thomas Lohse
3. Doc. Mgr. Jaroslav Bielčík Ph.D.

Tag der mündlichen Prüfung: 07.04.2021

Erklärung:

Hiermit erkläre ich, die Dissertation selbstständig und nur unter Verwendung der angegebenen Hilfen und Hilfsmittel angefertigt zu haben. Ich habe mich nicht anderwärts um einen Doktorgrad in dem Promotionsfach beworben und besitze keinen entsprechenden Doktorgrad. Die Promotionsordnung der Mathematisch-Naturwissenschaftlichen Fakultät, veröffentlicht im Amtlichen Mitteilungsblatt der Hum-boldt-Universität zu Berlin Nr. 42 am 11. Juli 2018,habe ich zur Kenntnis genommen.

I declare that I have completed the thesis independently using only the aids and tools specified. I have not applied for a doctor's degree in the doctoral subject elsewhere and do not hold a corresponding doctor's degree. I have taken due note of the Faculty of Mathematics and Natural Sciences PhD Regulations, published in the Official Gazette of Humboldt-Universität zu Berlin no. 42 on July 11 2018.

Berlin, 09.12.2020

Filip Nechanský

Acknowledgments

I would like to extend my thanks to the people who contributed in many ways to the completion of this thesis. First, I would like to thank Klaus Moening and Thorsten Kuhl for welcoming me at DESY Zeuthen as a PhD student, and for their support, supervision and for the freedom they provided me throughout the years.

Large gratitude goes to Timothée, who proved to be excellent colleague and friend and who patiently guided me during my work on the $t\bar{t}H(b\bar{b})$. I am also grateful to all the other members of the $t\bar{t}H(b\bar{b})$ analysis team, especially Alex, Nico, Manuel, Ian and Peter, for their help and cooperation on this difficult analysis. For his guidance during the $\gamma\gamma \to WW$ measurement I have to thank Oldřich, who continues to push me to always improve myself.

I have to thank Timothée and Thorsten again for their uneasy task of reviewing this thesis and also thank Martin for his valuable feedback on the text.

For the great time I had at DESY Zeuthen, I would also like to thank to, among others, fellow PhD students: Matthieu, Marianna, Martin, Ben and formerly also Baishali and Akanksha. My gratitude also goes to all the others who kept me busy and sane during the whole time, especially during the last months which were made difficult by the pandemic.

Finally, and most importantly, I have to thank to my family for their endless support and love during my studies. Without them I would not be able to get to this point in my life.

Abstract

The discovery of the Higgs Boson in 2012 confirms the Standard Model as the most successful theory describing the fundamental interactions of elemental particles. One of the important properties of the Higgs boson is its Yukawa coupling to the top quark, which in the Standard Model is the strongest due to the high mass of the quark.

This thesis reports on a measurement of the top-Yukawa coupling with data collected by the ATLAS detector from 2015 to 2018 at 13 TeV center of mass energy. The coupling is studied in $t\bar{t}H(b\bar{b})$ events, a final state containing decay products of two top quarks with additional emission of a Higgs boson, where the Higgs decays into a pair of bottom quarks. This decay channel of the Higgs Boson has the largest branching ratio, but is systematically limited by the description of the dominant background process $t\bar{t}b\bar{b}$, a $t\bar{t}$ with additional two b quarks in the final state.

The measurement takes advantage of the ability of the ATLAS detector to identify jets coming from a b quarks to construct analysis regions with various compositions of the signal and the background. To further separate the signal, a series of multi-variate algorithms is employed and the $t\bar{t}H$ process is then extracted using a profile likelihood fit.

The results are shown for the channel with a single lepton in the final state and for a combination with the dilepton channel. The background performance is studied in detail, where large mis-modeling is found. The measured ratio of the $t\bar{t}H$ production compared to the Standard Model prediction is found to be $\mu_{t\bar{t}H} = 0.84^{+0.45}_{-0.39}(\text{syst.}) \pm 0.21(\text{stat.})$. The result is in agreement with the Standard Model prediction and corresponds to an observed (expected) significance of 1.9σ (2.3σ), an improvement compared to the previous ATLAS measurement which reported 1.4σ (1.6σ).

Zusammenfassung

Die Entdeckung des Higgs-Bosons im Jahr 2012 bestätigt das Standardmodell als die erfolgreichste Theorie, die die grundlegenden Wechselwirkungen von Elementarteilchen beschreibt. Eine der wichtigen Eigenschaften des Higgs-Bosons ist seine Yukawa-Kopplung an das Top-Quark, die aufgrund der hohen Masse des Quarks im Standardmodell am stärksten ist.

Diese Arbeit berichtet über eine Messung der Top-Yukawa-Kopplung mit Daten, die vom ATLAS-Detektor von 2015 bis 2018 bei einem Massenschwerpunkt von 13 TeV aufgezeichnet wurden. Die Kopplung wird in $t\bar{t}H(b\bar{b})$-Ereignissen untersucht, einem Endzustand, der die Zerfälls-produkte von zwei Top-Quarks enthält und in dem zusätzlich ein Higgs-Boson emittiert wird, welches in Bottom-Quark-Paar zerfällt. Dieser Zerfallskanal des Higgs-Bosons hat das größte Verzweigungsverhältnis, wird jedoch durch die Beschreibung des dominanten Untergrundprozesses $t\bar{t}b\bar{b}$, ein Top-Quark-Paar mit zwei zusätzlichen b-Quarks im Endzustand systematisch beschränkt.

Die Messung nutzt die Fähigkeit des ATLAS-Detektors, Jets von einem b-Quark zu identifizieren, um Analysebereiche mit verschiedenen Zusammensetzungen von Signal und Untergrund zu konstruieren. Um das Signal weiter zu separieren, wird eine Reihe von multivariaten Algorithmen verwendet und der $t\bar{t}H$-Prozess wird unter Verwendung eines Profile-Likelihood-Fits extrahiert.

Die Ergebnisse werden für den Kanal mit einem einzelnen Lepton im Endzustand und für eine Kombination mit dem Dilepton-Kanal gezeigt. Die Untergrundgenauigkeit wird im Detail untersucht, wobei große Fehlmodellierungen festgestellt werden. Das gemessene Verhältnis der $t\bar{t}H$-Produktion zur Standardmodell-Vorhersage beträgt $\mu_{t\bar{t}H} = 0,84^{+0,45}_{-0,39}(\text{syst.}) \pm 0,21(\text{stat.})$. Das Ergebnis stimmt mit der Vorhersage des Standardmodells überein und entspricht einer beobachteten (erwarteten) Signifikanz von 1,9σ (2,3σ), eine Verbesserung gegenüber der vorherigen ATLAS-Messung, bei der eine Signifikanz von 1,4σ (1,6σ) ermittelt wurden.

Contents

CHAPTER 1

Introduction

The discovery of the Higgs Boson in 2012[1, 2] marks an important milestone in the confirmation of the Standard Model (SM). The story of the Higgs is, however, far from over. Though numerous main properties of this fundamental particle were determined since 2012 and were found to be in agreement with the theory[3], there is still a significant number of parameters which are yet to be measured. Will they agree with the Standard Model? Or will they finally give us clues into the physics that lies beyond it? These are the important questions physicist are facing today.

One of the big unknowns was, until recently, the strongest coupling related to the Higgs Boson: its coupling to the top quark. Why is measurement of the strongest coupling more difficult than of some other, weaker modes? The early measurements of the Higgs Boson were done in channels which have a relatively clear background, for example a decay to pair of Z bosons measured in a clear four lepton channel.

The challenge of a measurement of the coupling is the high mass of the top quark. The Higgs boson cannot decay into a pair of top quarks on mass shell and the cross-section of such process is strongly suppressed. That leaves only two possibilities to measure it. The first is through a process, which involves a quark loop. For example decay of the Higgs boson into a pair of photons. However, such measurement provides only an indirect evidence, since one does not know all the contributors to the loop and the result is strongly model dependent. Nevertheless, such measurement was performed and found to be in agreement with the prediction of the Standard Model[4].

That leaves only a single option for a direct measurement: a production mode. This is done in processes involving at least one top quark, which then radiates a Higgs boson. Example of such process would be the $t\bar{t}H$, a $t\bar{t}$ process with a Higgs radiating from one of the top quarks. It has the largest cross-section of all the potential processes used to study the top-Higgs Yukawa coupling and was successfully measured in 2018[5, 6]. Its properties were found to be in agreement with the standard model.

This thesis reports on a measurement of the $t\bar{t}H$ process where the Higgs boson further decays into a pair of b quarks. It was measured by the ATLAS experiment at center-of-mass energy 13 TeV. While the Higgs to $b\bar{b}$ is the primary decay channel of the Higgs with a 58% branching ratio[7], its contribution to the combined measurement is small. Though large in yield, the $t\bar{t}H(b\bar{b})$ suffers from low precision of the background estimation. The background is dominated by the $t\bar{t}b\bar{b}$, a $t\bar{t}$ process with an additional gluon in the final state which further splits into a pair of b quarks. The large number of jets in the final state complicates the modeling.

The $t\bar{t}H(b\bar{b})$ process is further split into additional channels based on the number of prompt leptons measured by the detector (either e or μ) coming from decays of the top quarks in the event. Single-lepton events are the primary focus of this thesis, though the combination with other channels is also discussed. Current iteration of the analysis is a continuation of previous analysis[8] with an extended data statistics.

The thesis starts with a general description of the theory behind the standard model and its simulation in chapters 2 and 3. The Large Hadron Collider (LHC), and ATLAS experiment are explored in the chapter 4 with more detailed description of particle the reconstruction and identification found in following chapter 5.

The $t\bar{t}H(b\bar{b})$ analysis in the single lepton channel is introduced in chapter 6 where general principles of the measurement are discussed. Background modeling, being the main challenge of the analysis, is presented as well. The analysis regions are defined, together with various techniques used in the preparation of input distributions. The chapter closes with a short introduction of the dominant systematic uncertainties.

The statistical analysis of the single lepton channel is then presented in chapter 7, large part of the chapter is dedicated to an investigation of impact of the limited statistics of Monte Carlo samples.

Since a major part of the analysis is done in combination with other channels, the chapter 8 is dedicated to their introduction and to description of various studies performed in the combination. Afterwards, interpretation of the results is presented.

Finally, chapter 9 summarizes and concludes the thesis. Possibilities of further extensions which could improve the $t\bar{t}H(b\bar{b})$ measurement are also discussed.

CHAPTER 2

The Standard model of particle physics

The goal of the measurement presented in this thesis is to study the $t\bar{t}H(b\bar{b})$ process and to provide a comparison to the Standard Model prediction. This chapter describes the fundamental properties of the SM, starting with the elementary particles in section 2.1. The theoretical framework responsible for the fundamental forces is described next in section 2.2, where the Higgs mechanism and Yukawa couplings responsible for the masses of SM particles are introduced.

Sections 2.5 and 2.6 focus on the theoretical description of properties of the top quark and the Higgs boson in the context of the SM, with some information specific to an LHC measurements. A short discussion on the production and decay channels used to study their coupling can be found in section 2.7.

2.1 Overview of the standard model

The SM currently represents the best understanding of the properties and interactions of elementary particles, in spite of several shortcomings discussed later in the chapter. It successfully describes three out of the four fundamental forces. The electromagnetic force was described historically first in the framework of the Quantum Electrodynamics (QED)[9], a result of several decades long effort to precisely describe interactions of fermions and photons. Later, QED was unified with the weak force under the electroweak (EW) theory[10–13]. The strong force was described latest in the framework of the Quantum Chromodynamics (QCD). Only the gravitational force is not included in the SM, since currently there is no proper unification of the SM and of the General Theory of Relativity (GTR). However, in the context of particle physics, the interaction strength of the gravitational force is significantly smaller than of the other forces and its possible impact is negligible.

The Standard Model itself is a quantum field theory, combining the EW and QCD in a single theory with the following non-abelian symmetry:

$$SU(3)_C \otimes SU(2)_L \otimes U(1)_Y, \tag{2.1}$$

where the C refers to the color charge, L is the left-handed coupling of the weak isospin doublet and Y denotes the weak hypercharge $Y = Q - T_3$, where Q is the EM charge in natural units and the T_3 is the third component of weak isospin[1]. The $SU(2)_L \otimes U(1)_Y$

[1]The weak isospin operator is defined as $\hat{T}_i = \tau_i/2$ where τ_i are the Pauli matrices.

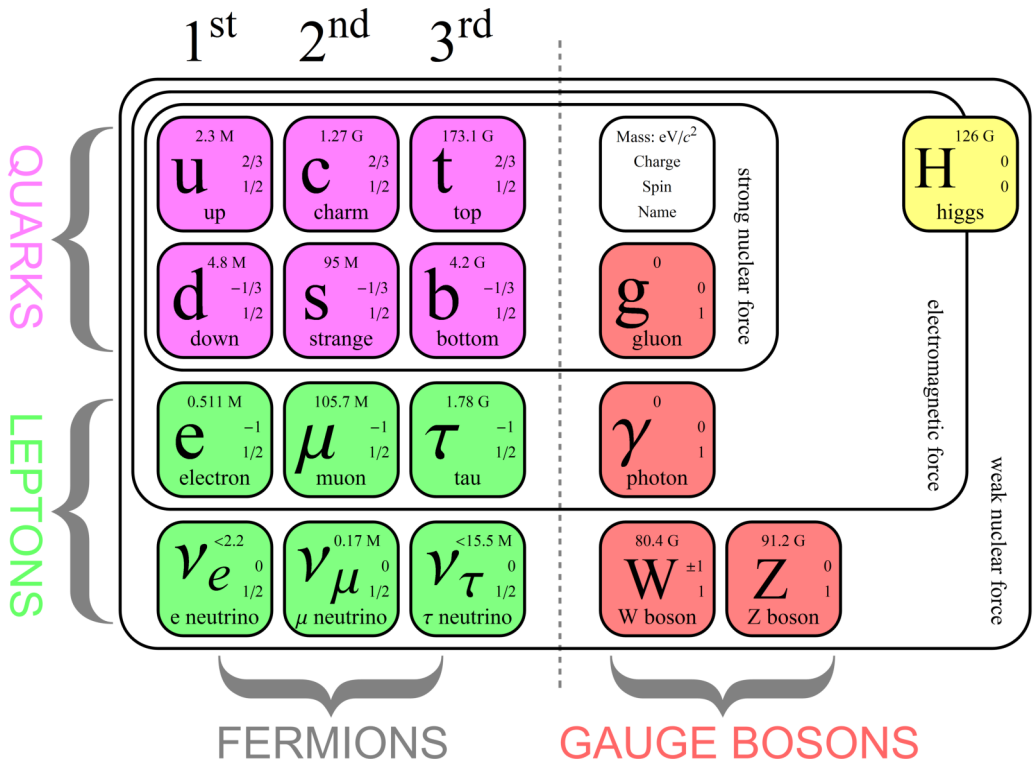


Figure 2.1: Particles of the Standard Model, and their main properties. Fermions are divided between quarks and leptons, and into three generations. The figure further specifies the type of the force they interact through (the large black box) and the carrier of the force (the gauge boson)[17].

symmetry of the group is broken through the Higgs mechanism to produce the particle masses[14–16].

The Lagrangian density of the Standard Model is:

$$\mathscr{L}_{SM} = \mathscr{L}_{EW} + \mathscr{L}_{QCD}, \tag{2.2}$$

where Lagrangian densities of the EW ($\mathscr{L}_{EW}$) and QCD ($\mathscr{L}_{QCD}$) will be described throughout the chapter.

Particles of the standard model and their main properties are summarized in figure 2.1. They can be divided into fermions and bosons. Fermions have a half-integer spin and are divided into three generations of particles. The first generation contains particles of which the stable observable matter, like atoms, is made up. Particles in the other generations except for neutrinos are heavier and less stable.

Fermions can be further divided between leptons and quarks. Leptons interact only through the electroweak force and are divided into charged particles and their corresponding neutrinos. The former have a charge of -1 and the particles of the three generations are called electron e, muon μ and tau τ, respectively. Neutrinos in the SM are massless and since they do not carry a charge, they do not interact electromagnetically. They are named based on the corresponding charged lepton, e.g. electron neutrino ν_e.

The second type of fermions are quarks. Compared to the leptons they carry the color charge and as such they interact through all three forces of the SM. There are six flavors of quarks in total, two for each generation. They can be divided into two groups based

on their charge Q, *up (u), charm (c)* and *top (t)* with $Q = 2/3$ and *down (d), strange (s)* and *bottom (b)* with $Q = -1/3$. Quarks naturally form bound states called hadrons. Pairs of quarks and anti-quarks create mesons, while three (anti-)quarks form (anti-)baryons. Bound states of four and five quarks also exist, called tetraquarks and pentaquarks.

The first category of bosons are force carriers and have a spin of size one. There is the massless photon γ, the carrier of the electromagnetic force. It interacts only with charged particles but does not have an electric charge itself. Similarly, the carrier of the strong force, the gluon g, is massless. Compared to the photon it holds a color charge and can self-interact. Finally, there are three heavy bosons $W^{\pm}$ and Z responsible for the weak interaction. The $W^{\pm}$ boson carries a charge of ± 1 while the Z is neutral.

The final particle of the SM is the Higgs Boson H, a scalar and an excitation of the Higgs field as a consequence of the broken symmetry. It interacts with massive particles: all particles with the exception of the photon, gluon and the neutrinos.

2.2 Electroweak theory

The electroweak theory is a result of unification of the QED and theory of weak interactions. This section offers a short summary of its historical development (for a more detailed overview see [18]). A large part of the section is dedicated to the electroweak symmetry breaking and introduction of the Higgs boson.

2.2.1 Quantum Electrodynamics

The Quantum Electrodynamics (QED)[9] is the first quantum field theory successfully describing electromagnetic interactions between charged fermions $\Psi(x)$ and photons, where photons are represented by a vector field $A_\mu(x)$. The QED is based on an Abelian gauge group $U(1)_Q$ and has the following Lagrangian density:

$$\mathscr{L}_{QED} = \bar{\Psi}(iD_\mu\gamma^\mu - m)\Psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu}, \tag{2.3}$$

where γ^μ are the Dirac matrices, D_μ is a covariant derivate $D_\mu = \partial_\mu + ieA_\mu$ and finally $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic field tensor. Introduction of the vector field is a direct result of the requirement on invariance under local gauge transformations of the $U(1)_Q$ symmetry. This principle of using local invariance to introduce vector fields into the model is one of the the fundamental principles behind the SM quantum field theories[18].

2.2.2 Weak interaction

Weak interactions were first observed in beta decays and first described by Fermi as a point-like interaction of the hadrons and leptons. In contrast to the QED, weak interaction violates the parity symmetry. Its lagrangian has the following form:

$$\mathscr{L}_{weak} = -\frac{G_F}{\sqrt{2}}[\bar{\Psi}_u\gamma^\mu(1-\gamma_5)\Psi_d][\bar{\Psi}_e\gamma^\mu(1-\gamma_5)\Psi_\nu], \tag{2.4}$$

where Ψ_x are the fermion fields of the interacting particles and $G_F \approx 1.12\times10^{-5}\mathrm{GeV}^{-2}$ is the Fermi constant. The parity violation is manifested in the Lagrangian in the $(1-\gamma_5)\Psi = \Psi_L$ terms, which only propagate to the left-handed fields Ψ_L.

The important milestone in the development of the electroweak theory was the idea to abandon the point-like structure of the interaction and instead have it mediated via

a heavy charged particle, the weak boson $W^{\pm}$. The introduction of the mediator was motivated by the fact that the point-like interaction violated the unitarity and lead to divergent cross-sections e.g. in $e - \nu$ scattering. Even though the W boson solved this issue, the model still had problems with other divergences and was eventually replaced by the united electroweak theory.

2.2.3 Electroweak unification

The electroweak theory successfully unifies the electromagnetic and weak interactions under a single $SU(2)_L \otimes U(1)_Y$ gauge theory. The left-handed lepton fields form a $SU(2)_L$ isospin doublet $L^{(l)} = \binom{\nu_L^l}{l_L}$, where the l is either e, μ or τ and the L subscript refers to the left-handed component Ψ_L mentioned previously (Ψ can be either l or ν^l). On the other hand the right-handed fields $l_R = (1 + \gamma_5)l$ form $SU(2)_L$ singlets.

A requirement on a local invariance of the non-abelian $SU(2)$ symmetry introduces a Yang-Mills field A^a (a=1,2,3) with a coupling g. Additional field B and coupling g' is a consequence of the abelian $U(1)_Y$ symmetry connected to the aforementioned hypercharge Y. The combination of the two fields introduces a new covariant derivative:

$$D_\mu^{EW} = \partial_\mu - (igA_\mu^a T_a) - ig' Y_x B_\mu, \tag{2.5}$$

where the second term only concerns the $SU(2)$ doublets. The value of Y_x depends on the type of field: $Y_L = -1/2$ and $Y_R^l = -1$. The resulting Lagrangian has the following form:

$$\mathscr{L}_{lepton}^{EW} = \bar{L}^{(l)}(iD_\mu^{EW}\gamma^\mu)L^{(l)} + \bar{l}_R(iD_\mu^{EW}\gamma^\mu)l_R. \tag{2.6}$$

The first two components of the field A can be associated with the W boson through the following relation: $W_\mu^{\pm} = \frac{1}{\sqrt{2}}(A_\mu^1 \mp A_\mu^2)$. Neither of the remaining fields A^3 and B corresponds directly to the electromagnetic field. Instead, it can be matched to a linear combination of the two fields by requirement on a purely vectorial interaction with the leptons and no interaction with the neutrinos. The complementary combination introduces a new neutral current field Z:

$$\begin{aligned} A_\mu &= \sin\theta_W A_\mu^3 + \cos\theta_W B_\mu \\ Z_\mu &= \cos\theta_W A_\mu^3 - \sin\theta_W B_\mu, \end{aligned}$$

where θ_W is the Weinberg angle related to the coupling constants of the two fields: $\tan\theta_W = \frac{g'}{g}$. The EM coupling constant is then $e = g\sin\theta_W$ and the Fermi constant can be matched with the coupling in the following way: $\frac{G_F}{\sqrt{2}} = g^2/8m_W^2$. The theory also relates the masses of the bosons through the Weinberg angle: $m_W = m_Z\cos\theta_W$.

In addition to the interaction with leptons, kinematic terms for the new gauge fields are required:

$$\begin{aligned} \mathscr{L}_{gauge} &= -\frac{1}{4}\vec{F}_{\mu\nu}\vec{F}^{\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \\ &= -\frac{1}{2}W_{\mu\nu}^{+}W_{-}^{\mu\nu} - \frac{1}{4}Z_{\mu\nu}Z^{\mu\nu} - \frac{1}{4}A_{\mu\nu}A^{\mu\nu} + \text{interactions between the bosons,} \end{aligned} \tag{2.7}$$

where $B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu$ (similarly to QED) and $F_{\mu\nu}^a = \partial_\mu A_\nu^a - \partial_\nu A_\mu^a + g\epsilon^{abc}A_\mu^b A_\mu^c$. The non-abelian nature of the $SU(2)$ is manifested in the last term. The ϵ^{abc} is three dimensional Levi-Civita symbol, which describes comutation relations between the isospin operators $[T_a, T_b] = i\epsilon_{abc}T_c$. The terms $W^{\mu\nu}, Z^{\mu\nu}$ and $A^{\mu\nu}$ are defined the same way as $B^{\mu\nu}$.

The eventual discovery of the neutral weak boson Z[19, 20] marked a huge success of the unified electroweak theory. In addition to describing interactions of bosons with leptons, it also introduces interactions between the bosons themselves. The γ naturally couples to the $W^{\pm}$ due to its charge, but in addition there is another three-point interaction ZWW. Finally, the theory introduced number of quartic couplings: $ZZWW$, $Z\gamma WW$, $\gamma\gamma WW$ and $WWWW$.

However, the theory still remains incomplete: there are still divergent amplitudes and it is not possible to introduce masses of the bosons and the fermions without breaking the gauge invariance.

2.2.4 Electroweak spontaneous symmetry breaking

The masses of the weak bosons are introduced through the Brout-Englert-Higgs mechanism. It is based on introduction of new degrees of freedom and subsequent spontaneous breaking of the $SU(2)_L \otimes U(1)_Y$ symmetry. The extra degrees of freedom are introduced as a complex weak iso-doublet Φ. The Higgs Lagrangian before the symmetry breaking is defined as:

$$\mathscr{L}_{Higgs}^{symm.} = (iD_{\mu}^{EW}\Phi)^{\dagger}(iD^{\mu EW}\Phi) + \mu^2\Phi^{\dagger}\Phi - \lambda(\Phi^{\dagger}\Phi)^2 \tag{2.8}$$

where the covariant derivative D_{μ}^{EW} is the same as in equation 2.5 with the hypercharge corresponding to the Higgs field is $Y_{\Phi} = 1/2$. The shape of the Higgs potential results in a degenerate minimum at $\Phi_0^{\dagger}\Phi_0 = -\mu^2/2\lambda = \frac{v^2}{2}$, if $\mu^2 > 0$. The symmetry is broken by choosing a specific ground state at $\frac{1}{\sqrt{2}}\binom{0}{v}$ and shifting the Φ potential appropriately. After removing extra degrees of freedom by an appropriate choice of the gauge (U-gauge as described in reference [18]), one arrives at a new form of the doublet Φ:

$$\Phi_U(x) = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}, \tag{2.9}$$

where $H(x)$ is the remaining degree of freedom, the Higgs scalar field. With the new field defined, the Higgs lagrangian takes on the following form:

$$\begin{aligned}\mathscr{L}_{Higgs} = &\underbrace{\frac{1}{2}\partial_{\mu}H\partial^{\mu}H}_{\text{Higgs kinem.}} - \underbrace{\lambda v^2 H^2}_{\text{Higgs mass}} \underbrace{- \lambda v H^3 - \frac{1}{4}\lambda H^4}_{\text{Higgs self-inter.}} \\ &+ \underbrace{\frac{1}{4}g^2v^2W_{\mu}^{-}W^{+\mu} + \frac{1}{8}(g^2+g'^2)v^2Z_{\mu}Z^{\mu}}_{\text{W/Z mass}} \\ &+ \underbrace{\frac{1}{8}(2vH+H^2)(2g^2W_{\mu}^{-}W^{+\mu} + (g^2+g'^2)Z_{\mu}Z^{\mu})}_{\text{H+W/Z inter.}}\end{aligned} \tag{2.10}$$

where the gauge fields A^a, B are once again interpreted as the $W^{\pm}$, and Z as described previously in section 2.2.3. The EM field A completely disappears, meaning that the photon remains massless. Following masses of the $W^{\pm}, Z$ and H are obtained:

$$\begin{aligned} m_W &= gv/2 \\ m_Z &= \sqrt{g^2+g'^2}v/2 \\ m_H &= \sqrt{2\lambda}v \end{aligned} \tag{2.11}$$

This in turn means that the value of v can be related to the Fermi constant $v = (G_F\sqrt{2})^{-1/2} \approx 246$ GeV. The value of the λ is, however, a priori unknown and the model does not predict the mass of the new particle.

2.2.5 Yukawa couplings and quarks in the EW theory

Lepton masses can be naturally included in the electroweak theory by introducing an $SU(2)$ invariant Yukawa interaction of the leptons and the Higgs field:

$$\begin{aligned}\mathscr{L}^{lepton}_{Yukawa} &= \sum_l -y_l \bar{L}^{(l)} \Phi_U l_R + h.c. \\ &= \sum_l \underbrace{m_l \bar{l} l}_{\text{lepton mass}} \underbrace{-\frac{y_l}{\sqrt{2}} \bar{l} l H}_{\text{lepton+H int.}}\end{aligned} \tag{2.12}$$

where beside the mass terms, interaction of leptons with the Higgs boson arises as well, with *Yukawa* coupling $y_l = \sqrt{2} m_l / v$ proportional to the lepton mass.

Quarks can be introduced into the EW theory similarly to leptons:

$$\mathscr{L}^{quark}_{Yukawa} = \sum_q \underbrace{-m_q \bar{q} q}_{\text{quark mass}} \underbrace{-\frac{y_q}{\sqrt{2}} \bar{q} q H}_{\text{quark+H int.}}$$

where $q = u, d, c, s, t, b$. The important difference with respect to the leptons is that for the down-type quarks the mass eigenstates q do not correspond to the weak eigenstates q_0 but instead are related through the unitary Cabbibo-Kobayashi-Maskawa (CKM) matrix:

$$\begin{pmatrix} d_0 \\ s_0 \\ u_0 \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \times \begin{pmatrix} d \\ s \\ u \end{pmatrix}. \tag{2.13}$$

The matrix can have a complex phase, resulting in CP violation. The weak eigenstates then appear in the interaction of the quarks with the weak bosons:

$$\mathscr{L}^{ew}_{quark} = \sum_{q=u,c,t} \bar{L}^{(q)}_0 (i D^{ew}_\mu \gamma^\mu) L^{(q)}_0 + \sum_{q=u,c,t,d,s,b} \bar{q}_{0R} (i D^{ew}_\mu \gamma^\mu) q_{0R}. \tag{2.14}$$

where the left-handed components have been grouped into three doublets $\binom{u_{0L}}{d_{0L}}$, $\binom{c_{0L}}{s_{0L}}$ and $\binom{t_{0L}}{b_{0L}}$. The corresponding hypercharges are $Y^{(q)}_L = -1/2$, $Y^{(q,up)}_R = 2/3$ and $Y^{(q,down)}_R = -1/3$.

This distinction between mass and weak eigenstates mainly modifies decays of the weak boson. If they were the same, there would be only three possible couplings with quarks: $W^+ \to u\bar{d}/c\bar{s}/t\bar{b}$. The mixing of the flavors leads to the possibility of all combinations of the up and down quark states. The amplitude of the coupling depends on the corresponding element of the CKM matrix V_{ij}. The largest mixing is between the d and the s quark (ratio of approx. 20:1), while for example the top decays almost exclusively to the bottom quark.

2.2.6 Complete electroweak Lagrangian

The Yukawa interaction terms are the last piece needed to construct the complete electroweak lagrangian density:

$$\mathscr{L}_{ew} = \mathscr{L}^{ew}_{quark} + \mathscr{L}^{ew}_{lepton} + \mathscr{L}_{gauge} + \mathscr{L}^{quark}_{yukawa} + \mathscr{L}^{lepton}_{yukawa} + \mathscr{L}_{higgs}$$

2.3 Quantum chromodynamics

Quantum Chromodynamics, the theory of the strong interaction, introduces a new $SU(3)_C$ color symmetry into the SM, represented by three color states a: red r, blue b and green g. To maintain the local symmetry, a new massless vector field is introduced in form of gluons. There are eight gluon fields G^i, related to the eight generators t^i of the $SU(3)$ group. The commutation relation of the generators $[t^i, t^j] = if_{ijk}t^k$ is defined through a structure constant f_{ijk}[21].

The covariant derivative of the QCD is defined as

$$D^{QCD}_{\mu,ab} = \partial_\mu - ig_s G^i_\mu t^i_{ab}, \tag{2.15}$$

where g_s is the strong coupling constant and the a, b indices refer to the color charge. The QCD Lagrangian density then reads

$$\mathscr{L}_{QCD} = \sum_q \bar{q}^a (iD^{ab,QCD}_\mu \gamma^\mu) q^b - \frac{1}{4} G^i_{\mu\nu} G^{i,\mu\nu}. \tag{2.16}$$

The non-abelian nature of the group symmetry gives a rise to three- and four-point self-interactions of the gluon field, since $G^i_{\mu\nu} = \partial_\mu G^i_\nu - \partial_\nu G^i_\mu + g f^{ijk} G^j_\mu G^k_\mu$.

Running coupling in the QCD

Modern computations of the SM rely on a perturbative expansion to produce predictions. This a priori produces infinities for finite order of the perturbation, which are removed through renormalization procedure. This introduces a dependence of the coupling $\alpha_s = g_s^2/4\pi$ on the renormalization scale μ_R, a *running* coupling[21]. Looking at the value of $\alpha_s(\mu_R^2)$ at energies close to the transferred energy Q (so $\mu_R = Q$) provides an effective value of the coupling in a given process[22]. At the leading order, the coupling takes on the following form:

$$\alpha_s(\mu_R) = \frac{1}{b_0 \ln(\mu_R/\Lambda)}, \; \Lambda \approx 200 \text{ MeV}, \tag{2.17}$$

where the $b_0 = (11n_C - 2n_f) = 21$ constant relies purely on the number of colors $n_C = 3$ and the number of quark flavors $n_f = 6$[2]. The Λ parameter refers to a scale where the perturbative theory in the QCD breaks down.

The behavior of the $\alpha_s(Q^2)$, displayed in figure 2.2, shows two main features of the strong coupling, the asymptotic freedom and confinement.

For large values of the interaction energy ($Q^2 \to \infty$), the strong coupling gets weaker and colored particles start to behave as if they were free. This is the principle of *asymptotic freedom* and it allows for better application of the perturbative theory at greater transfered energies Q^2, since the higher order terms have a smaller impact.

On the other side of the energy scale ($Q^2 \to 0$) the strong coupling increases significantly. This leads to a *confinement* at low energies and quarks start to form bound states: hadrons. During the formation of the hadron (hadronization), additional particles are produced. Because of this, high energetic quarks and gluons, produced for example in colliders, are manifested in the detector as collimated showers of hadrons, called jets. Because the values of the α_s in the low energy region are comparable to or larger than one, the perturbative expansion is no longer feasible.

Similarly to the QCD, one can also introduce running coupling of the EM coupling $\alpha_{em} = e/4\pi$. Because of the differences of the underlying theories, α_{em} decreases with

[2]For energies higher than mass of the top quark.

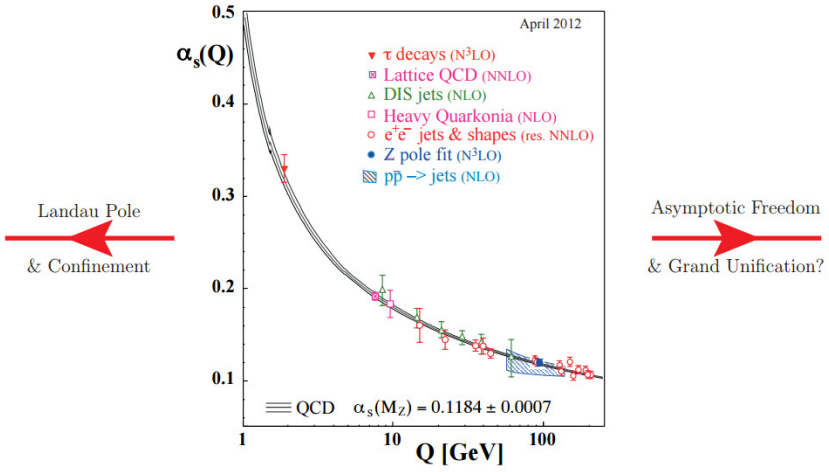


Figure 2.2: Dependence of coupling constant of the strong interaction α_s on Q^2. Taken from [22].

increasing distance and increases with energy. The strong and electromagnetic coupling will eventually become comparable at energies of order $10^{15} - 10^{17}$ GeV.

2.4 Limitations of the Standard Model

The Standard Model theory has been continuously tested in the past decades in numerous collider experiments, showing no deviation from predictions. Despite the tremendous success of the SM, measurements from cosmology and neutrino physics expose the theory to be incomplete, for example[21]:

- Observation of neutrino oscillations revealed that neutrinos are not massless. Though their mass can be introduced into the SM through the Higgs field, alternative mechanisms exist and the correct method and properties have to be yet established.
- There is a large asymmetry of the baryonic matter and anti-matter in the universe. However, the current level of CP violation present in the standard model is not large enough to explain the observed imbalance.
- The SM does not account for gravity, nor there is a successful theory merging it with the general relativity.
- Observations in cosmology reveal that large portion of the matter in the universe is represented by so called *dark matter*, which cannot be described in the SM.

2.5 Top quark

The heaviest particle of the Standard Model is the top quark. Its existence was theorized since the discovery of the τ lepton[23] and subsequent observation of b quark[24] to complete the third generation of quarks[3]. Observation of top quarks took some time because of

[3]The discovery of the last particle of the third generation, the tau neutrino, came much later in 2000[25].

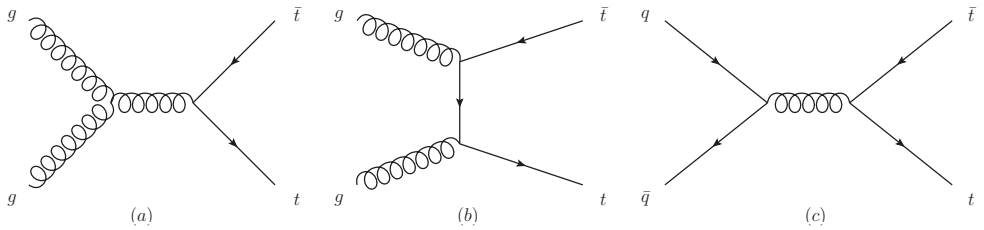


Figure 2.3: Tree-level Feynman diagrams of the three dominant $t\bar{t}$ production channels at the LHC, (a) and (b) with gluons in the initial state and (c) with quarks. Drawn using a JaxoDraw software package[28].

its high mass. Only after the masses of Z and W bosons were measured was the search narrowed to 170-180 GeV and the top was finally observed in 1995 at the Tevatron[26, 27].

Its high mass, at approximately 173.3 GeV[21], lends the top unique properties among the quarks. Since it is significantly heavier, it can decay into a real W boson, resulting in a short lifetime. Hence, it is the only quark which does not form bound states and decays immediately into bW final state (relevant CKM term $|V_{tb}| \approx 1$). Furthermore, the Yukawa coupling of the top quark to the Higgs boson is almost one $y_t = \sqrt{2}m_t/v \approx 1$, far greater than for any other quark. It makes it the primary contributor in e.g. quark loops connected to the Higgs.

At the LHC, the primary production channel of the top quark is top-pair production, also called $t\bar{t}$, with the three main production diagrams displayed in figure 2.3, where the diagram initiated by gluons generally dominates the production.

The top quarks decay almost always into bW, so the top-pair final state is divided mainly based on the products of the W decay. Either both Ws decay into leptons (≈10%), or one of them decays hadronicaly (≈45%), or both of them are hadronic (≈45%)[21].

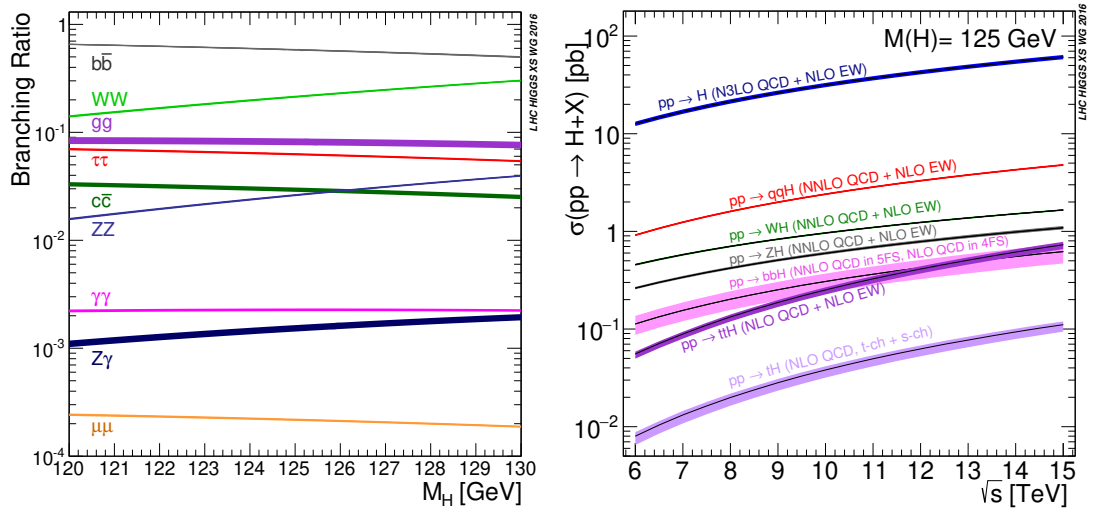
2.6 Higgs boson

The most elusive particle of the Standard Model was, until, recently the Higgs boson, discovered only in 2012[1, 2] with a mass at around 125 GeV. The measurements performed so far determined that it is a neutral C-even particle with spin 0, consistent with the SM Higgs boson. All decay and production channels measured so far also show no deviation from the theory[3].

The major decay channels of the Higgs boson are displayed in figure 2.4(a), where the decay into a pair of b quarks is the dominant final state ($BR \approx 58\%$). Though both the W and Z bosons have much larger masses, their production is suppressed since one of them has to be produced off-shell. Suppression is even larger for the top-pair final state, where neither of the quarks can be produced on-shell and the BR is negligible.

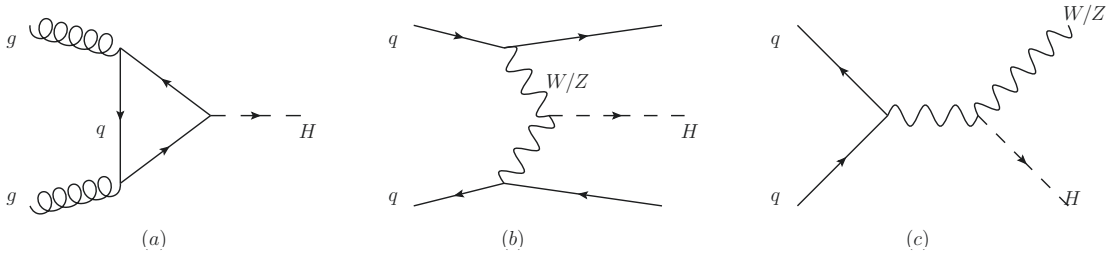
Various other particles contribute to the possible final states, where their cross-section reduces with a lower mass: τ, c, μ. Gluon and photons also appear among the dominant decay products, though they do not couple to Higgs directly. The decay is driven by a quark loop for gluons and a fermion-W loop for photons.

At the LHC, the Higgs boson is produced in several production channels. Their cross-section as a function of the center-of-mass energy $\sqrt{s}$ is displayed in figure 2.4(b). The dominant channel is gluon fusion induced through a quark loop, as displayed in the tree-level Feynman diagram in figure 2.5a). The top quark is the primary contributor in the loop.



(a) (b)

Figure 2.4: (a) Branching ratios of the dominant Higgs decays as a function of the Higgs mass m_H. (b) Main production channels of the Higgs boson as a function of the collision energy $\sqrt{s}$. Taken from [3].



(a) (b) (c)

Figure 2.5: Tree-level Feynman diagrams of the three dominant Higgs boson production channels at the LHC: (a) gluon fusion, (b) vector boson fusion and (c) WH/ZH associated production (Higgs-strahlung)[28].

The second leading process is the vector-boson fusion (VBF), where the Higgs is emitted from an intermediate W or Z boson of two quarks scattering in u or t channel (see figure 2.5b). A characteristic feature of this process is two forward jets[4] associated with the Higgs boson and particle activity in the central rapidity is suppressed. That allows to distinguish the VBF signal from the background, providing a method to measure Higgs coupling to the weak bosons. Higgs boson can be also produced in associated production with W and Z as shown if figure 2.5c).

Finally, the Higgs can also be studied in $t\bar{t}$ events where one of the top quarks emits the boson, a $t\bar{t}H$ process. As the subject of this thesis, it is discussed in more detail in the next section and generally throughout the document. Its cross-section is comparable to $b\bar{b}H$, but due to its clearer signature it is much easier to study.

2.7 Coupling of a top quark to Higgs boson

As the heaviest quark, the top has the largest Yukawa coupling to the Higgs boson. This makes it the dominant contributor in quark/fermion loops and its good understanding is critical to properly interpret some of the properties of the Higgs.

A priori, one can study any Higgs coupling in three major ways: Higgs decay, Higgs emission and from top quark's contribution to loops involving the Higgs. As was already discussed, the top quark is too heavy to be a product of Higgs decay. One can study it from the loops, where especially the gluon fusion production is convenient for this purpose due to its high cross-section. However, any such measurement has to make assumptions about the underlying theory. Particles not predicted by the SM can also contribute to the Higgs modes involving a loop or they can even modify the coupling of other particles to the Higgs.

This means that only Higgs emission from a top quark can offer a direct measurement of the top Yukawa coupling. The highest cross-section of such process comes from $t\bar{t}H$. Due to numerous decay channels of both the top quarks and the Higgs boson, the $t\bar{t}H$ process cannot be easily analyzed in a single analysis. In the ATLAS experiment the process is divided into several measurements, which are later combined.

There are four major channels in which $t\bar{t}H$ is measured. Currently, the most sensitive channel is $t\bar{t}H(\gamma\gamma)$ [29], where the two photons in the final state allow measure of the invariant mass of the Higgs with a good resolution, which makes it easier to measure. The analysis is, however, limited by low statistics. The $t\bar{t}H(ZZ)$ four lepton channel has similar properties but lower cross-section and is currently statistically limited[30]. If the Higgs produces at least one lepton (but excluding the four lepton channel), it is included in the $t\bar{t}H$ multi-lepton analysis[31, 32]. The biggest disadvantage of the multi-lepton channels is that not all decay products are detected, making it difficult to reconstruct the event kinematics.

Finally, if the Higgs boson decays into a pair of b-quarks, it is studied in the $t\bar{t}H(b\bar{b})$ analysis. It has the highest branching ratio and therefore highest statistics, but has other experimental shortcomings. The main challenge of the measurement comes from its dominant background: a $t\bar{t}$ process with additional two b-quarks in the final state. Modeling of events with additional heavy flavor quarks in the final state generally has large modeling uncertainties, which decreases the sensitivity of the measurement.

All four channels were combined into a single measurement[6], leading to an observation of the $t\bar{t}H$ process with 6.3σ significance. Similar analysis was performed by the CMS collaboration[5].

[4]Forward jet refers to a jet with $|\eta|$ >2.5, as described later in section 4.2.1.

CHAPTER 3

Event simulation

In order to compare the data to predictions of the Standard Model (or of some other theory), samples are simulated using the Monte Carlo (MC) method to represent the stochastic effects and the probabilistic nature of the underlying theory[21]. For this reason the simulated samples are often called Monte Carlo samples and the tools used to produce them are MC generators.

Their production is a complicated process, divided into two major steps. First, the underlying collision needs to be described, starting with the two protons in the initial state and their interaction and ending with stable particles in the final state. In the second stage, the propagation of particles from through the detector and simulation of detector response needs to be performed. The latter, being a detector specific process, is presented in the next chapter in the context of the ATLAS detector. Each collision and its subsequent products constitute a single event and the production of the MC samples can be called an event simulation or generation.

The simulation starts at small distances, where the colliding protons can be viewed as collections of partons (quarks and gluons). This is described in section 3.1. An interaction of these partons can lead to a hard-scattering, represented by a matrix element of the process presented in section 3.2. Higher orders of the QCD not present in the matrix element are approximated using parton shower models, introducing additional particles in the final state. For larger distances, colored particles are always bound: the bare partons have to undergo hadronization into color-less states. The parton shower and hadronization are described in section 3.3 and 3.4, respectively.

Finally, additional interactions between other partons of the two protons (*underlying event*) are possible and need to be simulated. This is discussed in section 3.5

3.1 Proton collisions and parton distribution functions

Protons are built out of two up and one down quarks, called *valence* quarks and accompanied by a *sea* of virtual quarks and gluons. Their contribution increases with increasing energy of the interaction. Proton collisions at high energies are then expressed as interactions of the constituent *valence* and *sea* partons.

To describe their interaction, only the parton type (quark or gluon) and its momentum is important. Since the transverse component of the momentum with respect to the incoming proton p is usually negligible, each parton is primarily described by the fraction of

the momentum of the proton it carries $x = p_z^{parton}/p_z^{proton}$, assuming the incoming proton travels in the z direction. The probability to find a parton with a fraction x is represented by a parton distribution function (PDF) $f(x, Q^2)$, where Q^2 is the energy scale of the interaction.

Currently, PDFs cannot be derived theoretically, since they involve low energy transfers which cannot be computed perturbatively due to high values of α_s (as described in section 2.3). Instead, they are measured in collisions of protons with various other particles in deep inelastic scattering.

Nevertheless, QCD allows to extrapolation of the PDFs from one energy scale to another through DGLAP[1] evolution equations[33–35] for quarks $q_i(x, Q^2) = f_i(x, Q^2)$ (where i refers to a quark flavor) and gluons $g(x, Q^2) = f_g(x, Q^2)$ within the proton:

$$\frac{\partial}{\partial \ln Q^2}\begin{pmatrix} q_i(x,Q^2) \\ g(x,Q^2) \end{pmatrix} = \frac{\alpha_s(Q^2)}{2\pi}\sum_j \int_x^1 \begin{pmatrix} P_{q_iq_j}(y) & P_{q_ig}(y) \\ P_{gq_j}(y) & P_{gg}(y) \end{pmatrix} \begin{pmatrix} q_j(x/y,Q^2) \\ g(x/y,Q^2) \end{pmatrix} \frac{dy}{y}, \tag{3.1}$$

where P_{ab} are Altarelli-Parisi splitting functions[35], which represent a probability of a parton b splitting to a parton a in the QCD computed at the leading order.

When describing a final state X in proton-proton collisions, its cross-section σ_X can be parameterized through a combination of the PDFs and cross-section $\sigma_{i,j\to X}$ of a hard scattering thanks to the *factorization theorem*, which separates the perturbative and non-perturbative components:

$$\sigma_X = \sum_{i,j} \int \sigma_{i,j\to X}(x_1p_1, x_2p_2, Q^2) f_i(x_1, Q^2) f_j(x_2, Q^2) dx_1 dx_2, \tag{3.2}$$

where p_1, p_2 are the momenta of the colliding protons and the formula sums over all possible production modes of the final state. The energy scale Q in the equation is usually referred to as the factorization scale μ_F.

An example of a PDF can be found in figure 3.1, derived in a combined measurement[36]. There are several notable features. The valence quarks dominate large values of the x, with contribution of the u twice as large as of the d. The sea partons diverge towards low values of x, with the gluon contribution an order of magnitude higher than that of quarks.

3.2 Hard interaction and matrix element

Usually, the goal of measurements at the ATLAS experiment[37] is to study properties of a specific process or group of processes, defined by their final state X. Due to the composite structure of the proton multiple production modes are possible. For example, as described in section 2.7, the $t\bar{t}H$ process can be produced from both pairs of gluons or quarks ($gg \to t\bar{t}H$ and $q\bar{q} \to t\bar{t}H$). Thanks to the factorization introduced previously, the description of the hard scattering can exclude the protons and be modeled only as an interaction of the component partons. The cross-section is then computed from the matrix element of the process.

Particles sensitive to the strong interaction tend to emit additional particles, either through gluon emissions, or gluons splitting into $q\bar{q}$ pairs. Hence, instead of producing exclusively a specific final state X, it is necessary to generalize it to $X + n$, where the n refers to extra partons in the final state. The inclusive matrix element is then simply a sum over the matrix elements of the possible final states.

[1]The acronym DGLAP refers to the names of its authors: Dokshitzer-Gribov-Lipatov-Altarelli-Parisi.

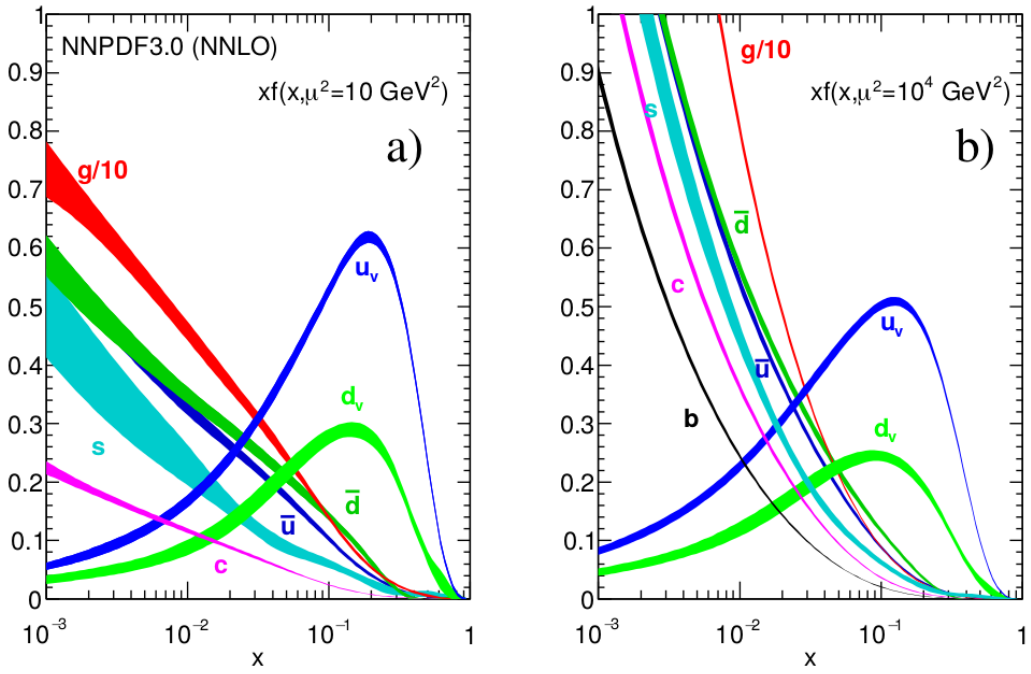


Figure 3.1: Examples of proton PDFs for several types of partons, shown for two factorization scales (a) $\mu^2 = 10$ GeV^2 and (b) 10^4 GeV^2[21]. Each curve represents $xf(x, \mu^2)$, where f is the PDF. These results are obtained in a global analysis designated NNLO NNPDF3.0[36].

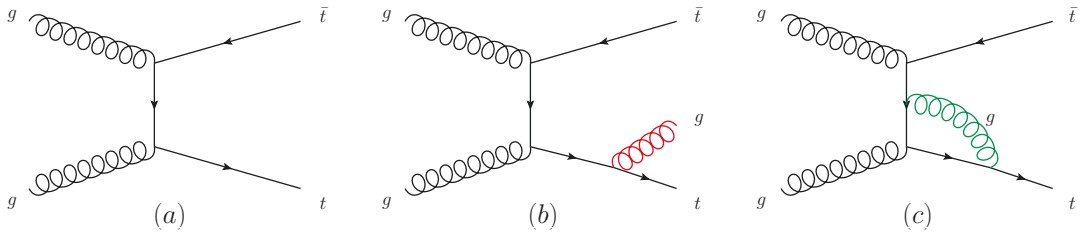


Figure 3.2: Feynman diagrams of the $t\bar{t}$ process in (a) the LO of QCD , (b) the next-to-leading order contribution with an additional emission (red) and (c) the virtual correction in the next-to-leading order involving a loop (green)[28].

Usually, the leading order (LO) describes the process of interest, as e.g. in figure 3.2(a). An additional radiation can be generally described by higher orders of the QCD (as in figure 3.2(b)). The radiation diagram is usually divergent and has to be accompanied by processes involving loops to cancel them out, as e.g. the one in figure 3.2(c). Matrix elements corresponding to higher orders of $\alpha_s(\mu_R^2)$, introduced in section 2.3, which for large energies has values smaller than one. This leads to their suppression and also results in a dependence on the renormalization scale μ_R^2.

3.3 Parton shower

Additional radiation can be described by the perturbative QCD framework up to the QCD scale Λ, where the value of α_s becomes larger than one. However, due to an increasing number of contributing processes, current generators usually provide only a matrix element at the leading order or the NLO. Parton showers offer a way to generate the additional soft (small momentum) and collinear (small angle) emissions, which dominate the production of additional particles. They are divided into the Initial State Radiation (ISR), connected to the initial parton from the proton before the matrix element, and Final State Radiation (FSR), which describes emission of the intermediate and final state particles, as illustrated in figure 3.3.

An example of Final State Radiation would be a gluon g emitted from a final state quark q under a small angle θ. The cross-section of such a process can be written as[21]:

$$\sigma_{qg} \approx \underbrace{\sigma_q}_{\text{bare cross-section}} \cdot \underbrace{\frac{\alpha_s(t)}{2\pi} P_{q.qg}(z)\mathrm{d}z \frac{\mathrm{d}\phi}{2\pi} \frac{\mathrm{d}\theta^2}{\theta^2}}_{\text{additional emission}}, \tag{3.3}$$

where z is the fraction of energy carried by the quark after the emission ($z = E_q/(E_q + E_g)$, where the E_q and E_g are energies of the quark and gluon after the splitting). ϕ describes the angle of the emission plane (plane defined by the two particles). Finally, $P_{q,qg}$ is the Altarelli-Parisi splitting function introduced already in section 3.1.

The energy scale of the splitting, usually called *hardness* t, can be defined simply as the *virtuality* of the splitting parton $t = p^2 \approx z(1-z)E^2\theta^2$ or based on the angle between the splitted partons θ $t = E^2\theta^2$, where E is the energy of the parton.

Additional emissions can be iteratively added in the same way as in equation 3.3, sequentially decreasing the hardness down to a cut-off value Λ, where the perturbative approximation breaks down. Since this results in a shower ordered by the hardness, one can have for example a *virtuality ordered* shower, which is used in the Pythia8 [38], or *angular ordered* shower[39], which is implemented in the Herwig generator[40].

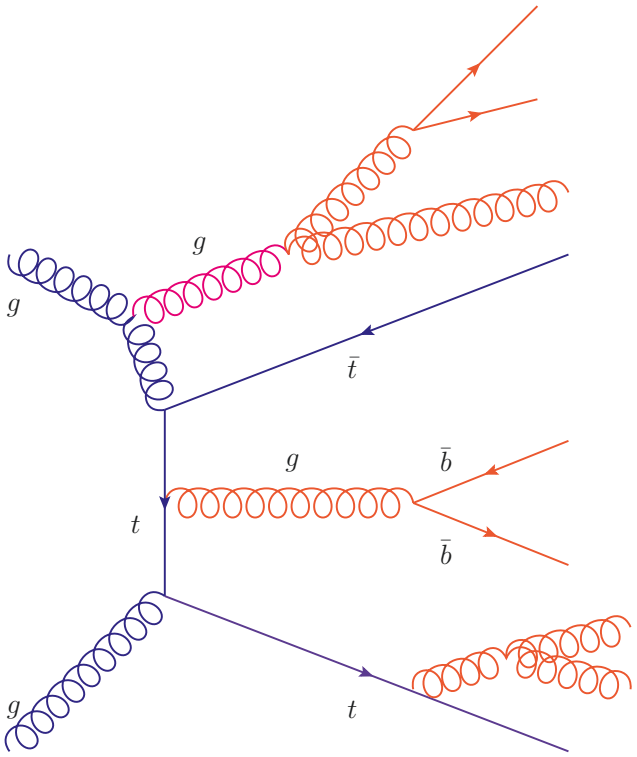


Figure 3.3: An example of a possible final state of the $t\bar{t}$ process, the blue color representing the underlying matrix element. The initial state radiation, displayed in pink, is connected to an initial gluon. The final state radiation, in orange, is emitted from the final state and intermediate top quark[28].

As mentioned in section 3.2, the amplitudes describing additional emissions are divergent and virtual loop corrections need to be included in the matrix element to compensate. These are, however, not part of the parton shower algorithms. Instead, the cut-off value Λ removes the collinear divergence $\theta \to 0$ and soft emissions $z \to 0, 1$. Similar precautions are applied to avoid soft emissions for the angular ordering $E^2\theta^2$, where one needs to introduce cut off values z_1, z_2 for emissions carrying a small fraction of the momentum, defining a selection $z_1 < z < z_2$.

Parton showers based on equation 3.3 modify the cross-section through the splitting functions and through the arbitrary value of the various parameters of the model (e.g. Λ, z_1, z_2). This dependence is removed by scaling the cross-section such that inclusively (when all possible final states are summed up) it maintains its original value σ_q. This is done by a Sudakov form-factor[21]:

$$\Delta_i(t, t') = \exp\left[-\int_t^{t'} \frac{\mathrm{d}q^2}{q^2} \frac{\alpha_s(q^2)}{2\pi} \sum_{j,k} P_{i,jk}(z)\mathrm{d}z\right] \tag{3.4}$$

which also describes a probability that particle i will undergo no emission between the two scales t, t'.

The parton shower is generated in the following way: for each particle i at a scale t, a uniform random number r between 0 and 1 is generated. One then finds a new scale t' by solving an equation $r = \Delta_i(t, t')$. If $t' > \Lambda$, a new emission is introduced at the scale t', otherwise the particle is considered to be a final state parton. This procedure is repeated until all partons are final state, producing new particles with a lower virtuality down to Λ.

A similar method is used for the Initial State Radiation[41], with a caveat that the initial partons have to be connected to the proton PDF. The cross-section is then modified in the following way:

$$\sigma_q(x)\mathrm{d}x f_q(x,t) \xrightarrow{\text{ISR}} \sigma_q(x)\mathrm{d}x f_q(x/z,t)\frac{\alpha_s(t)}{2\pi} P_{q.qg}(z)\mathrm{d}z\frac{\mathrm{d}\phi}{2\pi}\frac{\mathrm{d}\theta^2}{\theta^2}, \tag{3.5}$$

where the PDF changes based on the fraction of momentum z carried by the ISR. This transformation leads to a different form of the Sudakov form-factor as well:

$$\Delta_i^{ISR}(t, t') = \exp\left[-\int_t^{t'} \frac{\mathrm{d}q^2}{q^2} \frac{\alpha_s(q^2)}{2\pi} \sum_{j,k} \int_x^1 P_{j,ik}(z)\mathrm{d}z \frac{f_j(q^2, x/z)}{f_i(q^2, x)}\right], \tag{3.6}$$

which, compared to equation 3.4, includes the parton distribution function further modifying the emision probability.

Another alternative approach to the parton shower generation is so-called dipole shower[42], used by default in SHERPA [43]. It focuses on the soft emissions and considers quark pairs as dipoles instead of focusing on a single particle. It shares similarities to the hadronization models discussed in the next section, just in higher energy scales.

The methods for generation of parton showers introduced so far assumed zero mass of the quarks, which is reasonable for the u, d and s quarks, since $m_q \ll \Lambda$ and therefore their impact is negligible. That does not hold for the c, b and t quark, which due to their mass have lower probability of a collinear radiation[21]. This can be approximated through modification of the cut-off parameter Λ for heavy quarks[44], or more precisely through advanced methods described e.g. in references [44, 45].

NLO matching

Parton shower generators present only an approximation of the collinear and soft emissions and their description of hard emissions is inaccurate. Though numerous techniques to increase their precision exist, it is often advantageous to generate the hardest emission as a part of the matrix element.

The diagram with an NLO emission of a particle is, however, the same as an LO matrix element with additional emissions from the parton shower. There is an overlap between emission from the matrix element and the parton shower, which has to be removed to avoid double counting. Several methods exists, the most popular being POWHEG [46] and MC@NLO [47].

3.4 Hadronization

The event generation presented so far described the event evolution up to the energy scale Λ, where the perturbative QCD breaks down. However, the event still contains bare quarks and gluons, which need to be confined in hadrons. In the current paradigm, the *hadronization* is described by phenomenological models, which rely on some of the basic properties of the QCD.

String model

One of the most popular models, used for example in the PYTHIA8 generator, is the Lund string model[48, 49]. It is based on the fact that the potential energy between two quarks increases linearly with distance[2]: $V(r) = \kappa r$, where $\kappa \approx 0.2$ GeV^2.

Hence, the model pairs quarks of opposite color (based on the evolution of the parton shower) and connects them by one-dimensional *strings* representing the potential. As the event evolves, distances between the quarks grow, increasing the potential energy. Once the energy in the string is large enough, it breaks, creating a quark/anti-quark pair. This process is shown in figure 3.4.

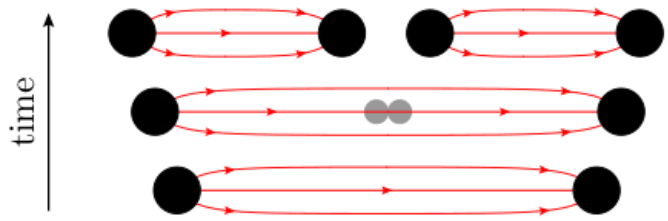


Figure 3.4: Evolution of a quark pair and subsequent breaking of the connecting string, leading to creation of a new quark pair[21].

Gluons are presented as "kinks" in the string connecting two quarks, introducing a transverse component to the originally one-dimensional object.

The strings break repeatedly until the energy of the quark pair is low enough to create a pair of mesons. To introduce baryons, the string can also break into a pair of diquarks, though the probability with respect to a pair of quarks is lower.

[2]This holds for distances larger than 1 fm, at shorter scales there is a Coulomb element $\propto 1/r$ as well.

Cluster model

Another popular hadronization scheme is the cluster model[50], used for example in HERWIG and SHERPA generators. It relies on the *preconfinement*[51], where the color structure of an event leads to creation of color singlet states from pairs of quarks.

First, gluons are forced to split into quark/anti-quark pairs (in contrast to the string model where they are part of the string). Pairs of quarks are then combined into color-less clusters. Clusters with a large energy ($E > 3$ GeV) are repeatedly broken into multiple clusters until all clusters have energy lower than 3 GeV.

Low mass clusters then form a single on-shell meson, where any excess momentum is distributed between nearby clusters to satisfy the conservation of momentum. The remaining clusters are treated as excited mesons, which subsequently isotropically decay into pairs of mesons or baryons.

Subsequent decays

The final state after the hadronization contains hadrons from the hadronization and leptons from the hard interaction. Some are unstable and their decay has to be evaluated. What constitutes "unstable" often depends on the measuring device, where the distinction mainly depends on whether the particle is able to reach the detector. For the ATLAS detector, described in the next chapter, hadrons are considered stable if $c\tau > 10$ mm and among leptons only the tau is considered unstable.

3.5 Underlying event and additional interactions

Beside the primary hard interaction which involves a large transfer of momentum and is modeled by the matrix element, additional interactions between other partons in the collision can take place. These extra processes, called an underlying event (UE), are softer but still produce new particles in the final state.

The UE is usually described as multiple parton interactions (MPI)[52]. They are dominated by QCD t-channel $2 \to 2$ exchanges of partons, with a differential cross-section proportional to $d\sigma \approx \frac{\mathrm{d}p_{\mathrm{T}}^2}{p_{\mathrm{T}}^4}$, where p_{T} is the transverse momentum of a final state parton (both partons have the same transverse momentum due to its conservation in head-on collisions). There is a divergence for soft $p_{\mathrm{T}} \to 0$ production, which is constrained by an introduction of a cut-off parameter p_{T}^{min}. It is motivated by the fact that for low p_{T} values the wavelength of the produced partons becomes larger than the distance of strong interaction, effectively leading to a screening[21]. It is usually set to a value close to the size of the QCD perturbative cut-off $\Lambda \approx 0.2$ GeV.

The average number of $2 \to 2$ interactions $\langle n \rangle$ in a proton-proton collision is then described by the following equation[21]:

$$\sigma_{2\to 2} = \langle n \rangle \cdot \sigma_{tot}, \tag{3.7}$$

where σ_{tot} is the total cross-section of proton-proton interactions.

The actual number of generated MPIs in an event is then based on the Poisson distribution with a mean value set to $\langle n \rangle$. An additional suppression of the MPI is based on the available momentum in the parent proton: the sum of the fractions of the momentum carried by the partons cannot exceed 1.

3.6 Overview of Monte Carlo generators

Several MC event generators are used in the analysis presented in this thesis. Some generators mainly provide the NLO matrix element and perform the necessary matching of the parton shower. The primary generator of this type used in the analysis is PowhegBox [46, 53], which uses Powheg for the matching. As an alternative, MadGraph5_aMC@NLO [47] generator is utilized. It uses MC@NLO to match the shower.

To produce the parton shower and perform the subsequent hadronization, two *multipurpose*[3] generators are used. The default generator used in the ATLAS experiment is Pythia8 [38]. It uses a virtuality ordered shower and the Lund string model for the hadronization. The multipurpose generator Herwig [40] is used as an alternative model. It has an angle-ordered shower and depends on the cluster model for the hadronization.

The last generator used in the analysis is Sherpa [43], which provides an NLO matrix element and also produces the parton shower and performs the hadronization. It contains an NLO matrix element while providing the subsequent parton shower in the dipole formalism. By default, it uses the cluster model for hadronization.

All generators contain several parameters which affect the generated events. These parameters are often derived, or tuned, from data. Specific set of parameters is then called a tune.

[3]Multipurpose generators are able to simulate the whole event, including the matrix element, parton shower and hadronization. However, they usually include only an LO matrix element.

CHAPTER 4

The ATLAS experiment

The analysis presented in this thesis was performed using the ATLAS detector[37], one of the main experiments at the Large Hadron Collider (LHC)[54], an accelerator colliding protons and heavy ions at record energies and luminosity. The experiment is used to study long predicted physics phenomena and to search for completely unknown processes.

This chapter starts with a description of the LHC in section 4.1, followed by section 4.2 with a presentation of the ATLAS experiment, including aspects like the trigger and data-taking. Section 4.3 then closes with the simulation of the detector.

4.1 The Large Hadron Collider

The LHC[54] is located at CERN[55], a laboratory for particle physics situated near Geneva on the border between France and Switzerland. It is a part of the CERN accelerator complex[56], a series of accelerators producing particles of various energies to numerous experiments. Though both protons and heavy ions are collided at the LHC, most of the following discussion focuses on the former as this thesis reports on a measurement in proton-proton collisions.

The whole accelerator complex, displayed in figure 4.1, starts with a simple hydrogen bottle. By stripping electrons of the atoms, protons are produced and sped up to kinetic energy of 50 MeV through the first accelerator Linac 2. The Proton Synchrotron Booster (PSB) and the Proton Synchrotron (PS) accelerate the particles further up to 1.4 GeV and 25 GeV respectively. Right before the LHC is the Super Proton Synchrotron (SPS), where the protons reach energies of 450 GeV.

Finally, the LHC itself can accelerate protons up to the design energy of 7 TeV, though at the time of the writing of this thesis only 6.5 GeV energy has been achieved. The collider itself is situated in a circular tunnel with circumference of 27 km, buried deep underground in order to minimize the background from the cosmic radiation.

Protons at the LHC are accelerated by radiofrequency cavities through a pair of vacuum tubes, particles in each going in the opposite direction. They are not accelerated individually, but in bunches of around 10^{11} protons. During collisions there are over 2000 bunches in the LHC, separated by 25 ns and bended by a strong 8T magnetic field, created by a series of superconducting magnets cooled down to almost 0K.

Bunches at the LHC intersect in four interaction points, where the four large experiments are situated. Beside ATLAS, it is ALICE [58], which primarily studies heavy ion collision,

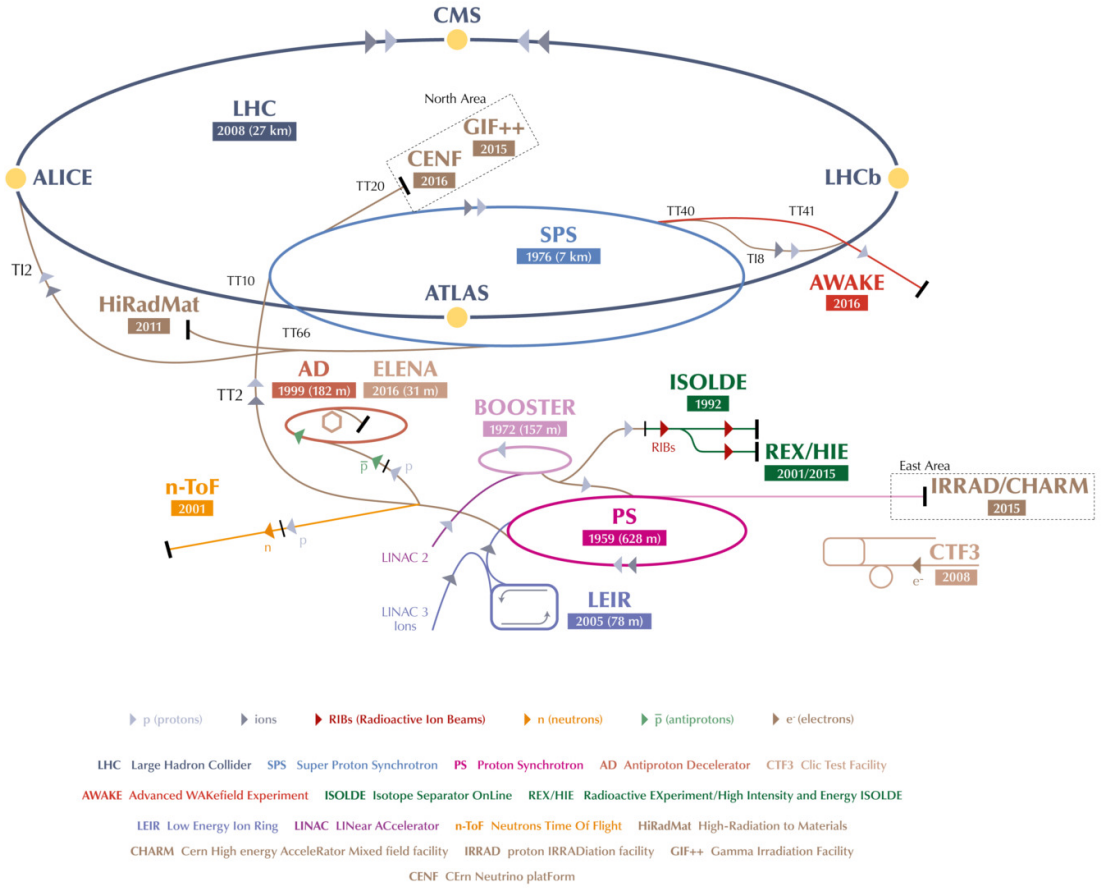


Figure 4.1: Graphic showing the series of accelerators and colliders which are part of the CERN accelerator complex and the path the protons or ions take through it (displayed through the grey arrowheads). Adapted from [57].

LHCb[59], focusing mainly on B-hadron physics, and finally CMS[60]. Both the ATLAS and the CMS experiments are general purpose detectors, built to study the variety of interactions produced in the collisions of the LHC.

4.1.1 Luminosity

Bunches of protons at the LHC collide with a frequency of 40MHz at the four interaction points. The rate of interactions and of various proton-proton processes is, however, not trivially related to the collision rate.

For any given process X, its rate R_X in collisions at the LHC is simply a product of two variables, the cross-section of the process σ_X (already introduced in chapter 3) and the instantaneous luminosity L:

$$R_X = L\sigma_X. \tag{4.1}$$

The cross-section represents the underlying physics and depends only on the center-of-mass energy of the collisions, while the luminosity encompasses all the factors coming from the properties of the collider. By integrating over the data-taking time, one gets the expected total number of events N_X as $N_X = \sigma_X \int L dt$, where $\int L dt$ is simply called integrated luminosity.

The instantaneous luminosity can be determined through the following formula[61]:

$$L = \frac{f_r n_1 n_2}{2\pi \Sigma_x \Sigma_y}, \tag{4.2}$$

where f_r is the LHC revolution frequency (40MHz) and n_1, n_2 refer to the number of particles in the two colliding bunches (around 10^{11}). The two parameters Σ_x, Σ_y represent the convoluted beam sizes in the vertical and horizontal direction. They are not easily related to a single property, but are convolutions of various factors like the shape of the bunches and their collision angle. In practice, the properties necessary to measure luminosity are determined using a van der Meer scan technique[62], performed in specialized LHC data-taking runs. More details about luminosity measurement can be found in references [61, 63].

The design luminosity of the LHC in proton-proton collisions for the purpose of the ATLAS and CMS experiments was $L = 1 \cdot 10^{34}\text{cm}^{-2}\text{s}^{-1}$[54], but due to the excellent performance and various improvements to the machine this value was surpassed, reaching peak luminosity of $1.9{\times}10^{34}\text{cm}^{-2}\text{s}^{-1}$ during 2018[61].

Luminosity provided by the LHC is not constant. As the bunches collide repeatedly, the number of protons in each bunch decreases and with that also the luminosity (see formula 4.2). Furthermore, the beam parameters can change during the year.

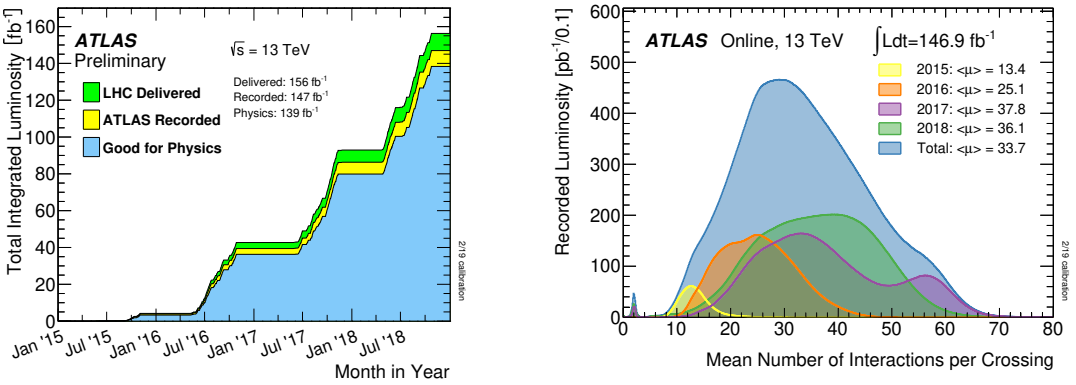
Pile-up

Larger luminosity means larger statistics to analyze, but it also comes with difficulties. Since the LHC is not colliding individual protons but bunches, more than one collision per bunch crossing can take place. The actual number of interactions is a random Poisson variable with a mean value μ. It can be determined through formula[63]:

$$\mu = L\sigma_{inel}/f_r \tag{4.3}$$

where $\sigma_{inel} \approx 80\text{mb}$ is the combined inelastic cross-section[61].

From equation 4.3 it simply follows that with a larger luminosity directly comes a larger number of protons interacting per bunch crossing. This leads to larger noise and background in the detector, and to a more difficult reconstruction of particles.



(a) (b)

Figure 4.2: (a) Evolution of the integrated luminosity over the Run-2 data taking period, shown for the luminosity provided by the LHC, the luminosity actually recorded by the experiment and finally the integrated luminosity usable for most physics measurements. (b) Distribution of the average number of interactions in different years of the Run-2[64].

Generally, there is one hard-scatter event per bunch crossing, producing an event interesting for a physics analysis. The other interacting protons usually result in a soft (low-energy) scattering. Only hard-scatter events are considered as a signal and the additional interaction in the same bunch crossing is considered as background[1], which plays a significant role in the reconstruction of objects. Henceforth, they will be called *pile-up* interactions.

4.1.2 LHC and ATLAS data-taking

The data taken by the ATLAS experiment so far can be classified into two main periods. First, there is the Run-1, a data-taking period from between years 2009-2012, where the center of mass energy was up to 8 TeV. After that, there was the Long Shutdown 1 (2012-2015), a period during which no collisions took place and several upgrades where added to both the accelerator and the detectors.

The analysis presented in this thesis relies on the Run-2 data, a sample provided by the LHC at $\sqrt{s} = 13$ TeV in years 2015-2018, with an integrated luminosity at around 156fb^{-1}. Due to various factors, for example the availability and the performance of the detector, the actual luminosity usable for most physics analyses is lower: 139fb^{-1}with an uncertainty of 1.7%[61]. The evolution of the integrated luminosity over the years can be seen in Figure 4.2(a). Currently, the LHC has another down period (Long Shutdown 2), which will continue at least until 2021.

The mean number of interactions during the Run-2 can be found in figure 4.2(b), showing an increase over the first few years as the collider was ramping up. The average number of interactions over the whole 4 years is found to be 33.7[64].

[1]It is important to note, that there are analyses which specifically study soft-interactions.

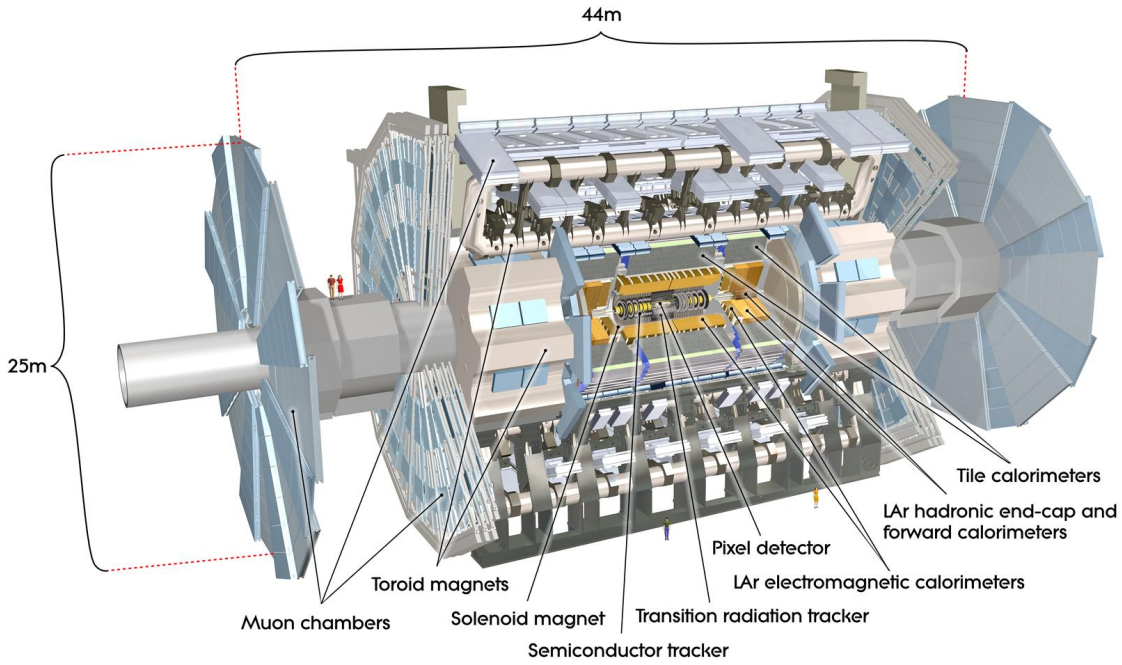


Figure 4.3: Cut-away of the ATLAS experiment showing the major parts of the detector[37].

4.2 The ATLAS detector

The ATLAS detector is one of the two general purpose detectors at the LHC. Its goal is to measure and search for diverse SM and BSM physics processes in the high energy and high luminosity proton-proton and heavy ion collisions.

The detector itself is positioned in the Insertion Point 1 of the LHC. Symmetric around the interaction point, the ATLAS is 25m high and 44m long cylindrical detector weighing approximately 7000 tonnes. It is a collection of sub-detectors, magnets and supporting infrastructure designed to reconstruct products of the collisions while withstanding the harsh radiation environment of the LHC.

The whole detector is displayed in figure 4.3, showing the main parts and features of the detector. The innermost part of the ATLAS is the Inner Detector (ID), a combination of three tracking detectors used to reconstruct tracks of charged particles and their interaction vertices. The entire ID is submerged in an approximately 2T solenoid magnetic field which bends the tracks and thus allows measurement of their transverse momentum and charge.

Around the ID is the calorimeter system. Closest to the ID is the electromagnetic (EM) calorimeter, which primarily measures photons and electrons, while hadrons are mainly measured in the hadronic calorimeters, which stop all the remaining particles except for muons and neutrinos. Those, due to their low interaction rate, pass through the whole detector. The outermost tracking detector, the Muon Spectometer (MS), is surrounded by a toroidal magnetic field and allows an extension of the measurement of muon tracks.

The various components of the detector are generally divided into two main categories based on their geometry. First, there is the barrel part, which has cylindrical structure centered around the beam-pipe close to the interaction point. In larger distances from the interaction point are the end-caps, which have a planar circular geometry perpendicular to the beam-pipe.

4.2.1 ATLAS coordinates and variables

The center point of the detector, where the collisions take place, is also the center of all major coordinate systems. The z direction of the Cartesian coordinate system then corresponds to the direction of the beam-pipe and the x-y plane is transverse to it, with the x axis pointing towards the center of the LHC and the y axis pointing upwards.

Variables computed in the transverse plane play a large role in the ATLAS measurements. This is for two main reasons. First, since the protons collide head-on the transverse momentum $p_{\mathrm{T}} = \sqrt{p_{\mathrm{x}}^2 + p_{\mathrm{y}}^2}$ of all particles produced in the proton-proton interaction has to sum up to zero. Second, the solenoidal magnetic field of the inner detector allows to measure the p_{T} of charged particles directly from the curvature of the track.

Aside from the p_{T}, two other properties defined in the transverse plane are used: the transverse energy $E_t = \sqrt{m^2 + p_T^2}$ and missing transverse momentum $E_T^{miss} = -|\sum_i \vec{p_{\mathrm{T}}}^i|$. The latter uses the requirement on the zero-sum of $\vec{p_{\mathrm{T}}} = \vec{p_{\mathrm{x}}} + \vec{p_{\mathrm{y}}}$ and is based on the sum of the p_{T} of all measured particles in a given interaction vertex. The missing momentum then corresponds to particles not detected by the detector, mainly neutrinos.

In the polar coordinate system, the azimuthal angle ϕ is computed in the transverse plane, while the polar angle θ is computed with respect to the beam axis. However, in context of HEP using θ is not always practical due to large dependence of the number of produced particles on the angle, where most of them are produced closer to the beam-line.

The pseudorapidity η, defined as $\eta = -\log\tan(\theta/2)$, is often used instead. Based on the detector geometry discussed later, one refers to small $|\eta|$ values as the *central* region and large values of $|\eta|$ as *forward*. The actual border value strongly depends on the context, but for purpose of this thesis it will be mainly based on the coverage of the Inner Detector where the central region refers to values of $|\eta| < 2.5$.

The pseudorapidity is a low mass approximation of the rapidity $y = 1/2\log[(E + p_{\mathrm{z}})/(E - p_{\mathrm{z}})]$. The distance in a $\phi - \eta$ ($\phi - y$) plane, defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ ($\Delta R_y = \sqrt{\Delta y^2 + \Delta\phi^2}$) is used when talking about distances between various objects.

Due to the shape of the detector a cylindrical coordinate system is used as well. It is defined using the ϕ, z mentioned previously and r, which simply denotes the distance from the center in the transverse plane. Another important set of variables is defined in the context of the particle tracks. Due to a finite resolution of the detector, the particle originating in the collision vertex are reconstructed with some finite distance from it. This distance is parameterized through impact parameters (IPs). To define them, a point of the closest approach of the track to the vertex in the transverse plane is found. The distance between this point and a vertex in the $(r - \phi)$ plane is called a transverse impact parameter d_0 and the longitudinal distance is denoted simply z_0. Their graphical representation can be found in figure 4.4.

4.2.2 The Inner Detector

The main purpose of the ID is to reconstruct tracks of charged particles produced in the collisions of the LHC. Up to thousands are produced in every collision[37], and since the ID sits closest to the interaction point, it has to have a good granularity and a great resistance to radiation damage. It is around 6 meters long and 2 meters high and covers a pseudorapidity range up to 2.5. It consists of three sub-systems. The innermost parts are the pixel and the Semiconductor Tracker (SCT), silicon based detectors with a high granularity and radiation resistance. In larger radii, where the conditions are less harsh, is the Transition Radiation Tracker (TRT). All three systems are divided into the barrel and end-cap parts. The overall structure of the ID can be seen in figure 4.5.

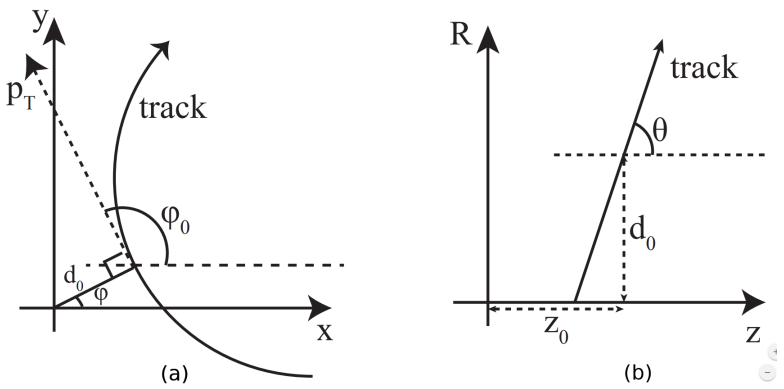


Figure 4.4: Definition of the track impact parameters[65]. First, the point of the closest approach of the track to associated vertex in the transverse plane is found. Then, its distance in transverse plane (a) is denoted d_0 while its longitudinal distance is called z_0 (b).

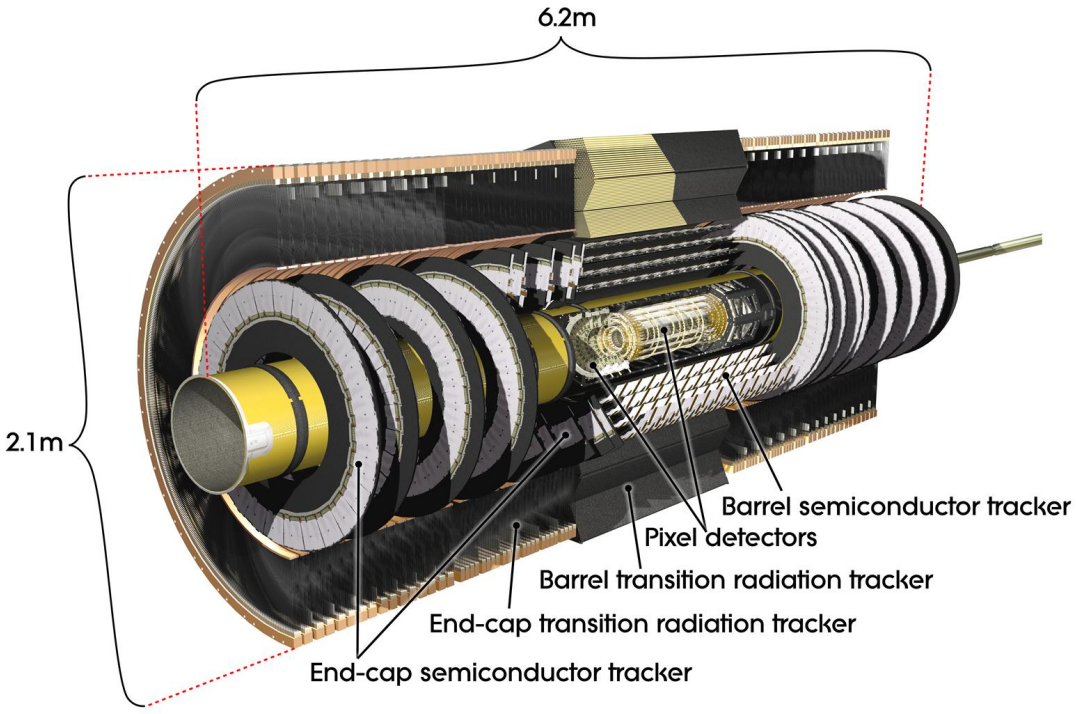


Figure 4.5: Cut-away of the ATLAS Inner Detector showing the major parts except for the IBL, which was added to the device during the Long Shutdown 1[37].

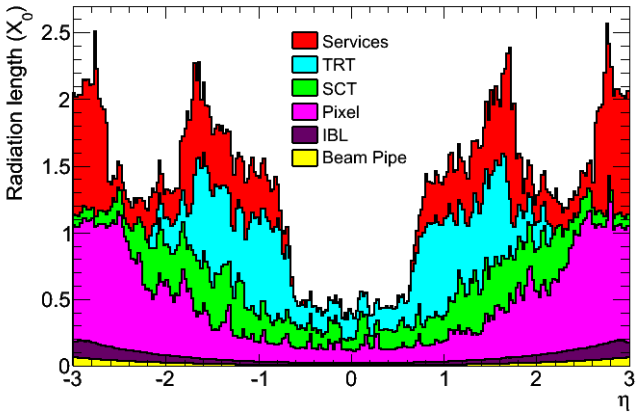


Figure 4.6: Amount of the material of the Inner Detector in units of radiation length as a function of pseudorapidity, divided between detector's components[66].

The reconstruction of the momentum of charged particles relies on their behavior in a magnetic field. Unobstructed, they would follow a helical path, but can change their trajectory due to interactions with the detector. In order to minimize effects of the scattering, the amount of material in the tracking detector has to be as low as possible. The amount is usually represented using a radiation length X_0, which describes a distance over which electron loses 1/e of its energy. For the ID the material is found to be below 0.5 radiation lengths for the center of the detector ($|\eta|$ <0.5) and goes up almost 2.5 for larger values of $|\eta|$. The amount of material as a function of η can be found in figure 4.6.

Pixel detector

The pixel detector is the first of the two silicon semiconductor detectors. These rely on a creation of electron-hole pairs in the volume of the semiconductor by the passing charged particle. The created charge is then collected and measured.

The pixel detector itself consists of two parts, the original three layer system[37], present in the ATLAS experiment from the beginning of the data-taking, and the Insertable B-layer (IBL), an extension of the original system placed closer to the beam-pipe[66]. It was added to the detector during the Long Shutdown 1 to improve its granularity and to mitigate negative effects of the radiation damage in the other layers of the Pixel detector. It also leads to a better performance of the identification of jets from B hadrons.

The IBL has only the barrel geometry and is placed approximately 3.3 cm from the center of the beam-pipe. It consists of modules, each covered by pixels only 50x250 μm^2 large. The spatial resolution of the IBL is found to be 10 μm in the $r-\phi$ plane and 66.5 μm in the z direction[67].

The original three layer system of the pixel detector has both a barrel and end-caps, each divided into three layers. The barrel sits between 50 and 122 mm from the beam-pipe, while the discs are between 495 to 650 mm from the center on each side of the detector. Both parts are covered by the same sensors with pixels of minimal size 50x400μm^2, almost twice as big in the z direction as those in the IBL. For this reason the resolution in barrel is basically the same in the $r-\phi$ plane (10 μm) but worse in the z direction (115 μm)[68].

The whole pixel system covers pseudorapidity $|\eta| < 2.5$ and has 80 milion readout channels, majority of the total readout channels in the ATLAS detector.

Semiconductor Tracker

The second silicon semiconductor detector, the SCT, is placed around the pixel system. Its barrel part is situated between 30 and 51 cm away from the beam-pipe while the end-caps sit at $84 < z < 274$cm. The barrel has only 4 layers, while the end-caps have 9 discs on each side of the detector. Covering this area with pixels would increase the number of read-out channels significantly, complicating the data collection. To maintain the number of chanels to a reasonable value without notably degrading the resolution, a strip layout is used instead.

Each layer of the SCT is covered with strips 80 μm wide and 12 cm long. To determine the exact position of a particle, two back-to-back stereo layers of sensors are placed together with a small 40 mrad angle between the strips. It determines the position of a track with a resolution of 17 μm perpendicular to the strip and 580 μm in the parallel direction[68]. The SCT, though larger than the pixel, ends up with only around 6 million channels, while covering the same region of pseudorapidity $|\eta| < 2.5$.

Transition Radiation Tracker

The outermost part of the ID, the TRT, does not rely on a silicon technology as the previous two sub-detectors. Instead, it is a gaseous detector constructed out of 4mm wide straw tubes filled by default with a mixture of gases (70% Xe, 27% CO_2, 3% O_2)[37]. A wire goes through the middle of the tube and a constant voltage is applied between it and the outer wall. A particle passing through the mixture ionizes the gas, generating a current.

The TRT is divided into a barrel and end-caps, with straws 144 and 37 cm long respectively. Unlike the silicon sub-detectors, it only covers $|\eta| < 2$ and it only allows identification of the position in the $r - \phi$ (barrel) or $r - z$ (end-cap) direction with a resolution of 130 μm. On the other hand, it offers 73 (160) layers of straws in the barrel (end-cap). It has around 350 thousand read-out channels.

In addition to a simple detection of a passing particle, the TRT allows for an electron identification through transition radiation, generated in a polypropylene mixture placed between the straws. The emission of the transition radiation depends on the relativistic factor γ, which for given energy of the particle is high for the electron due to its low mass.

However, due to leakage of the gas mixture, some tubes had to be refilled. Due to budget constraints, this was done with an Argon instead of Xenon dominated mixture. As a result, the electron-discriminating properties of the TRT are suppressed.

4.2.3 Calorimeters

The philosophy behind calorimeters is almost an opposite to that of tracking detectors like the ID. Instead of simply measuring points along the particle trajectory with as little material as possible, calorimeters simply stop the particle in a large chunk of material. The energy deposited by the particle and by the products of its interaction with the material is collected and measured. This allows determination of the position and energy of the original particle.

Calorimeters used in the ATLAS are composed of two alternating parts, an absorber which stops the particle and a sampling material in which its energy is measured. Each calorimeter unit then consists of numerous layers of absorbers and samplers. The calorimeter system is further divided between two main parts, the electromagnetic (EM) calorimeter, which stops and detects mainly electrons and photons, and the hadronic calorimeter, which measures the energy of hadrons. In contrast to the ID, the calorimeters are able to also

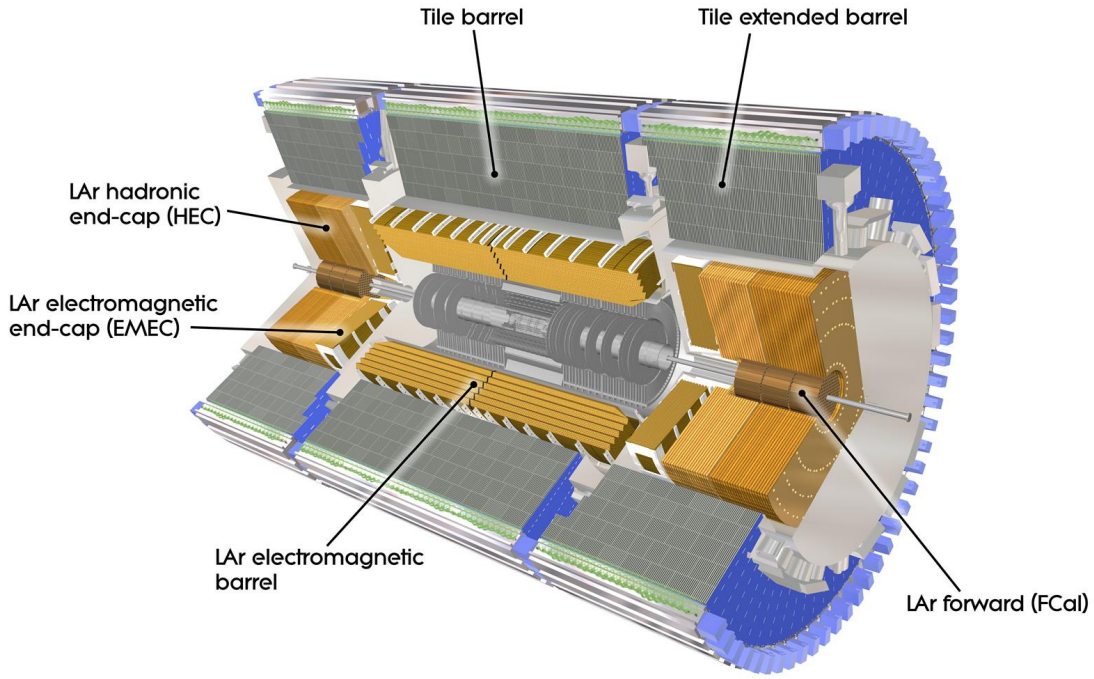


Figure 4.7: Cut-away of the ATLAS calorimeter system[37].

detect neutral particles. Out of all SM particles, only the neutrinos and muons have the interaction rate small enough to pass the calorimeters, though muons still deposit a small amount of energy.

The ATLAS calorimeter system is displayed in figure 4.7, showing its further division beyond the simple EM and hadronic components. This complex system covers a large $|\eta| < 4.9$ range, far beyond the range of the ID. Large coverage and ability to stop most of the particles allows measurement of the missing energy E_T^{miss} with a good precision[37].

Electromagnetic calorimeter

Out of the two calorimeters, the EM calorimeter is the one closer to the beam-pipe and has a higher granularity to better measure photon and electron showers. It is divided between a barrel and end-caps and overall covers absolute pseudorapidity up to 3.2. This coverage is assured by a large number of cells with a high granularity in the $\eta - \phi$ plane, with size as low as 0.025×0.025. The detector has an accordion geometry, which provides a full ϕ coverage. The EM calorimeter uses lead as the absorber and liquid argon (LAr) as the active material. Electrodes then collect the charge generated in the LAr by the particle shower. The whole sub-detector has over 173 thousand read-out channels.

Electrons and photons generally produce additional electrons and photons in interactions with a material, resulting in an EM shower, displayed in figure 4.8. Most commonly, electrons emit a photon via bremsstrahlung, while photons split into an electron-positron pair (photon conversion), which can in turn undergo bremsstrahlung. The distance in which an electron looses 1/e of its energy due to bremsstrahlung is called a radiation length X_0. It strongly depends on the type of material the electron is traversing and is a practical unit to measure absorption properties of an EM calorimeter. The thickness of the EM

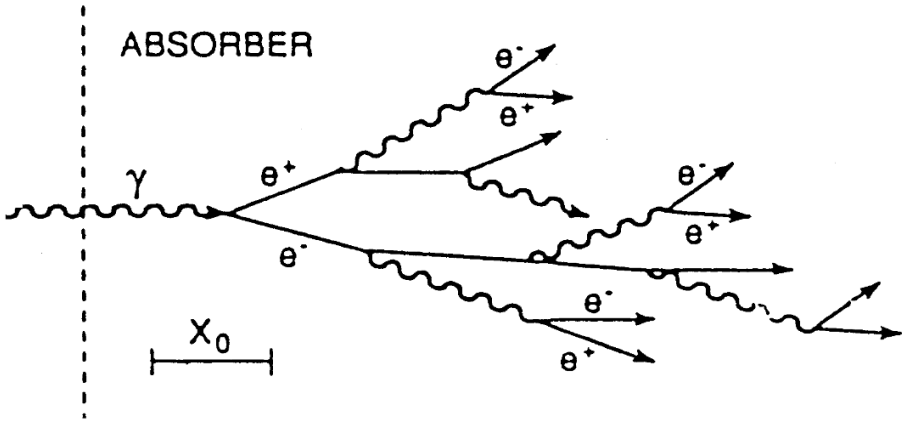


Figure 4.8: An illustration of an electromagnetic shower in an absorber[69].

calorimeter is over $22X_0$ in the barrel and over $24X_0$ in the end-caps, large enough to contain most of the EM shower.

Hadronic calorimeter

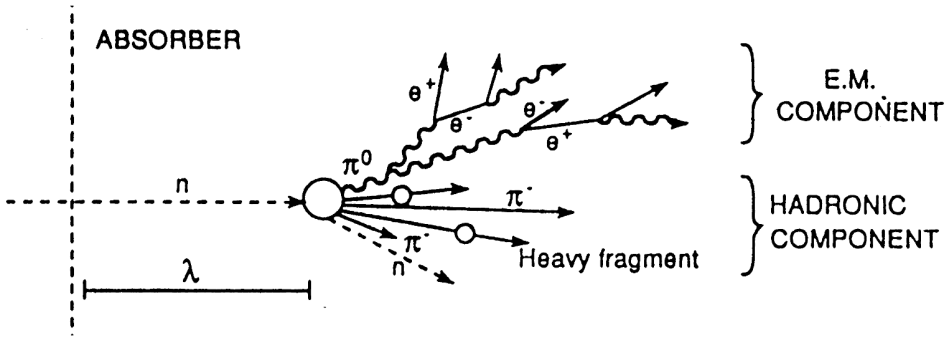


Figure 4.9: An illustration of a hadronic shower in an absorber[69].

The hadronic calorimeter aims to stop all the remaining particles (with exception of muons and neutrinos) within its volume. This means mainly hadrons, which in the first step lose energy mainly through nuclear interactions with the absorber though charged particles also lose energy by ionizing the detector material. This produces photons and electron, resulting eventually in EM showers. This is illustrated in figure 4.9. The distance over which hadrons lose on average 1/e of their energy due to nuclear interactions is called a nuclear interaction length λ_I. It offers a good description of absorption properties of the detector. The hadronic calorimeter is divided between three sub-detectors:

The **Tile calorimeter** forms an outer cylindrical envelope around the other calorimeters with combined coverage of the barrel and end-cap up to $|\eta| = 1.7$, segmented into 64 parts in the azimuthal direction. It has a granularity of 0.1×0.1 in the $\eta - \phi$ plane, a significantly larger cell compared to the EM calorimeter. It uses steel as the absorber and plastic scintillators for the sampling, both forming several layers (tiles) with a thickness

ratio between the steel and plastic around 4.7:1. Together they form a volume with an average length of $7.4\lambda_I$, covering most of the particle shower. The light in the scintillator is collected and measured using wave-shifters and photomultipliers located at the edges of the tiles.

The first extension of the hadronic calorimeter to higher $|\eta|$ is the **LAr hadronic end-cap calorimeter (HEC)**, covering region of $1.5 < |\eta| < 3.2$, overlapping slightly with the Tile calorimeter. Both end-caps are further divided into four layers, each constructed out of 32 wedge-shape modules. The absorber is made out of copper of thickness 25-50 mm, interlaced with 8 mm gaps filled with LAr. The granularity of the HEC in the $\eta - \phi$ plane is the same as that of the Tile calorimeter for low pseudorapidity (0.1×0.1) and slightly larger (0.2×0.2) for the forward parts of the detector.

The highest values of the $|\eta|$ are covered by the **LAr forward calorimeter (FCal)**, placed at $3.1 < |\eta| < 4.9$. It has a lower granularity then the other hadronic calorimeters with cells of size between 3.0×2.6 and 5.4×4.7 in $\eta - \phi$. It is constructed out of three layers, the first using copper as the absorber while the latter two use tungsten. It uses LAr as the active medium. Overall the FCal has depth of 10 λ_I.

All three sub-detectors of the hadronic calorimeter together have around 19 thousand read-out channels.

4.2.4 Muon spectrometer

The outermost part of the ATLAS detector, the Muon spectrometer, is responsible for a detection and precise measurement of muons. This is ensured by a system of tracking detectors embedded in a magnetic field generated by three toroidal magnets. One is responsible for the field in the barrel while the other two provide it in the end-caps.

The tracking systems itself is composed out of four sub-systems, two of which are responsible for high precision tracking, while the other two function as triggers (described later in section 4.2.6). An overview of the muons system can be found in figure 4.10.

Monitored Drift Tube

The Monitored Drift Tube (MDT) chambers form the largest part of the muon system and cover range of $|\eta| < 2.7$, similar to that of the ID. Their main purpose is a precision tracking, achieved with use of three to eight layers of drift tubes 3 cm in diameter, leading to an average resolution of 80μm per tube and 35 μm per chamber. They are filled with a gas mixture of Ar/CO_3 (93/7) maintained at a pressure of 3 bar[37]. The voltage between the wire and the outer wall is set to around 3000 volts.

The MDT chambers are divided between a barrel and end-caps but, in contrast to e.g. the TRT, the drift tubes are oriented in the ϕ direction, in both the barrel and the end-caps.

Cathode Strip Chambers

The expected particle multiplicity at high $|\eta|$ is too high for the MDT. For this reason it is substituted with Cathode Strip Chambers (CSC), a multiwire proportional chamber detector with cathodes divided into strips, each side perpendicular to the other. This leads to a good spacial resolution of 40 μm.

The CSC has an end-cap geometry with eight segments covering the whole ϕ range and $2 < |\eta| < 2.5$, with wires of the chamber being perpendicular to the beam-pipe. Each chamber has four layers, thus giving four independent measurements of the track trajectory.

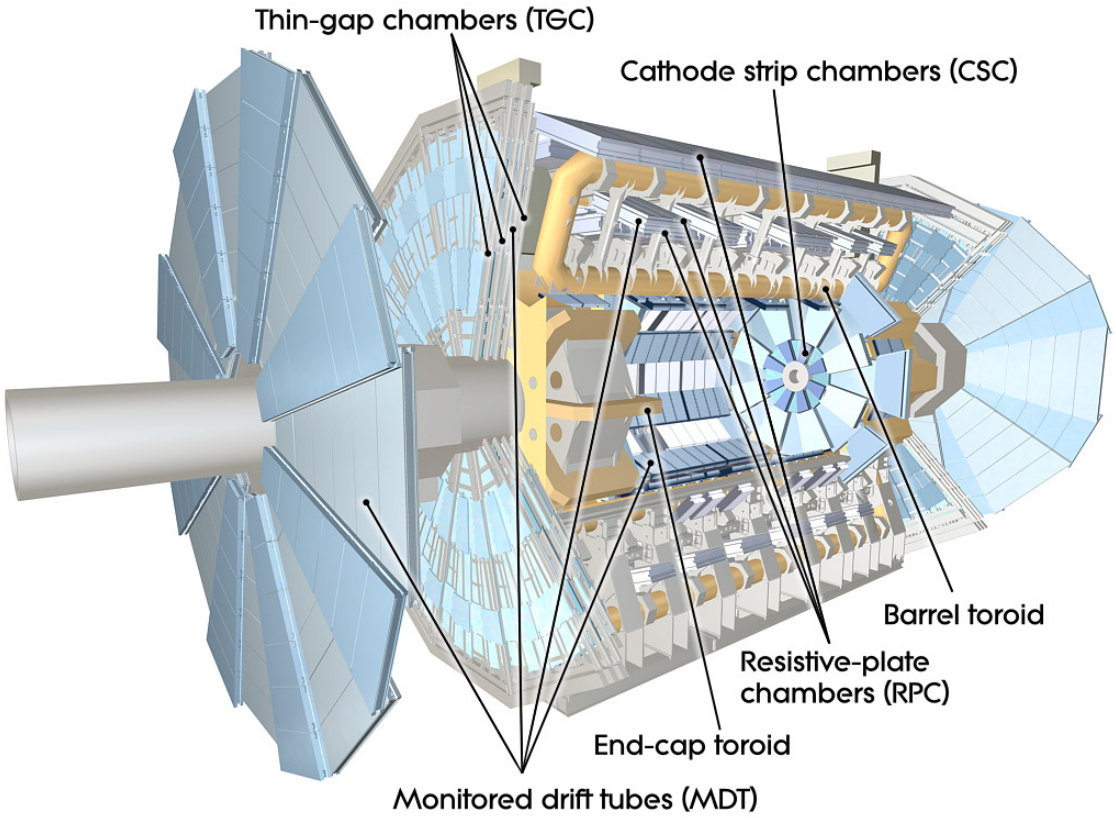


Figure 4.10: Cut-away of the ATLAS muon system[37].

Trigger chambers

The main purpose of the remaining sub-detectors, the Resistive Plate Chambers (RPC) and the Thin Gap Chambers (TGC), is to provide a fast (15-25 ns) information about the presence of muons to the trigger system. They also provide additional measured coordinates for a track reconstruction. They cover the whole ϕ range and $|\eta| < 2.4$.

The barrel is covered by the RPC, implemented within the barrel of the MDT in three layers covering up to $|\eta| = 2.5$. It is a gaseous detector built of two parallel resistive plates 2 mm apart under high voltage. Particles traversing the detector produce avalanches in the volume of the detector and the charge is then collected on the plates. The plates are segmented into strips to allow for a position measurement.

Larger values of $|\eta|$ are covered by the TGC. It complements the MDT end-caps and is used for the muon triggers in a pseudorapidity range $1.05 < |\eta| < 2.4$. It is a proportional multi-wire chamber, the radial coordinate determined by the wires and the azimuthal by radial strips.

4.2.5 Forwards detectors

The detector systems covered so far could be also called central detectors, which measure relatively small values of $|\eta|$. They are used for most of the ATLAS analyses. There is, however, an extensive program focusing on forward (high $|\eta|$) physics. Example would be measurement of dissociation in photon-induced processes[70] or luminosity measurements[61, 63].

There were three forward detectors in operation during the Run 1, the LUCID[71], the ALFA[72] and the ZDC[73], situated in various positions along the beam-pipe as displayed in figure 4.11. During the LS1, additional forward detector, the AFP[74] was added at around 200 meters away from the interaction point.

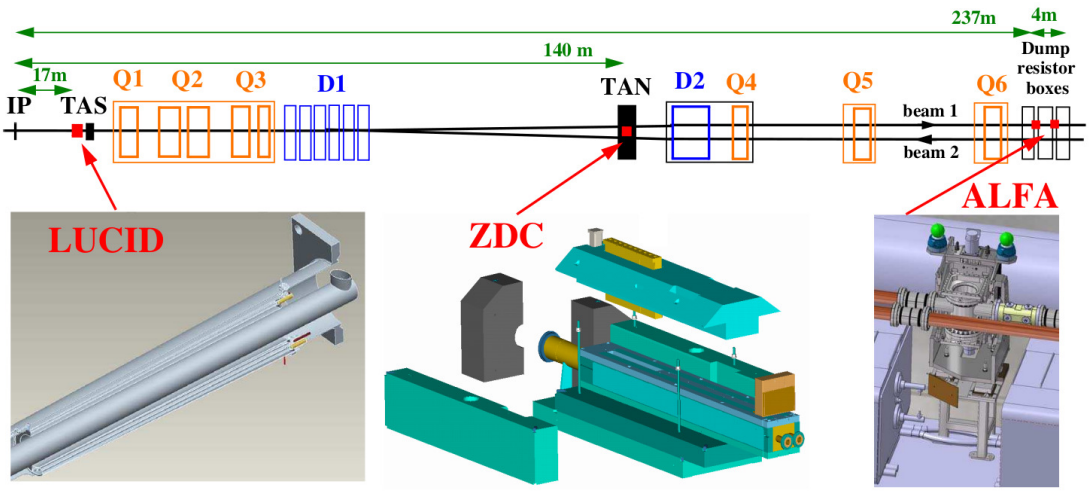
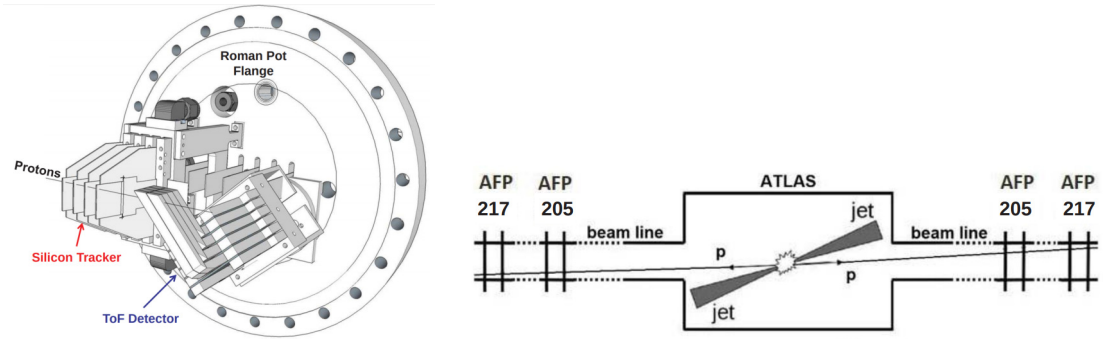


Figure 4.11: Placement and visualization of the ATLAS forward detectors in function during the Run-1 of the ATLAS data taking[37].

The LUCID is the first of the forward detectors and the one closest to the interaction point. It is a Cherenkov detector which measures part of the products of inelastic $p - p$ scattering. Its main primary purpose is luminosity measurement, which works under



(a) (b)

Figure 4.12: (a) A scheme showing the main components of the AFP detector and (b) its position along the beam-pipe (in meters)[75].

assumption that the number of charged particles going through the detector is proportional to number of interactions per bunch crossing, as was already discussed in section 4.1.1

The luminosity can be also measured through elastic scattering, but such scattering results in end products (protons) being deflected in a smaller angle on average (3μrad)[37]. A detector measuring these products has to be as close to the beam-pipe and as far from the interaction point as possible. The detector designed for this purpose is ALFA, a Roman-pot detector which puts the measuring element at 1 mm distance from the beam[72].

Between the LUCID and ALFA sits the ZDC. It is a calorimeter, whose main purpose is to measure forward neutrons in heavy-ion collisions. This is important in order to determine a centrality[2], a critical variable of the heavy ion physics program.

The **ATLAS Forward Proton (AFP)** detector offers a possibility for a more detailed measurement of the forward protons. This allows study of elastic and diffractive scattering or dissociative components of photon-induced processes[74]. The detector itself, displayed in figure 4.12, is a combination of a silicon tracker (similar to the IBL), used to reconstruct proton tracks, and a time-of-flight detector, important to reject background from pile-up.

4.2.6 Trigger system and data acquisition

The amount of data produced in collisions of the LHC would be practically impossible to record[3]. Furthermore, most of these events would be irrelevant to the main physics goals of the Large Hadron Collider (LHC). For this reason, a series of requirements, called triggers, is placed on individual events in order to determine whether a given event is recorded.

The trigger system of the ATLAS detector is divided into multiple levels[77]. First, there is the Level-1 (L1), a hardware trigger which reduces the event rate from 40MHz down to 100kHz. It predominantly decides based on the information from the muon triggers and the calorimeters, though other devices can be used to trigger, like the ZDC or the LUCID detector. The measured signal is processed by the Central Trigger Processer (CTP), which then decides whether the event is collected or not. It also applies a preventive dead time to avoid overlapping read out or to keep the front-end buffer from overflowing.

The L1 is followed up by the high-level trigger (HLT)[78], a software based trigger which reduces the rate down to 1kHz. Its decision is based on a Region-of-Interest (ROI) information from the L1 triggers, though compared to it some basic reconstruction is

[2]Centrality describes the impact parameter between the centers of the colliding ions.

[3]The average size of ATLAS events usable for physics is 1MB[76]. If all events were to be recorded, the output rate would be 40TB per second.

performed as well. This includes a reconstruction of tracks of charged particles or a first rough computation of the missing E_{T}. Even b-jet (jets from a decay of B hadron) identification is performed in this step. It is, however, important to note that this is only a crude fast reconstruction and more precise algorithms are used in the offline step, which will be discussed more in the next chapter.

Both levels of the trigger can be summarized by the Trigger menu, which defines all the different configurations of the L1 and HLT objects which determine whether a given event is selected or not.

The selection on objects is often accompanied by requirements on the quality of the reconstructions, which together with the kinematic requirements determine the average event rate. In case the event rate turns out to be too high, a pre-scale n can be applied, where only one in n events is recorded. This allows recording of events with a looser selection but most of them are discarded randomly.

In addition to the triggers, other requirements are put on each event to assure quality events for data analysis. For example, the detector has to be fully functional and there usually has to be a reconstructed primary vertex (explained later in section 5.2.2).

4.3 ATLAS simulation

The previous chapter provided a description of Monte Carlo simulation, which generates events expected in the collisions at the LHC. In order to compare the data to the theory, one then has two options. One can try to compensate for the effects of the detector, like resolution and efficiencies. This is called unfolding and it comes with numerous caveats, which do not allow for straightforward application for any distribution of interest.

An easier way is to directly compare the data to the theory. It requires a simulation of the propagation of the event through the detector and then uses the same procedure to select events as for the data. This is the purpose of the ATLAS simulation[79], a framework based on the GEANT software package[80]. It requires a good understanding of the detector response and geometry in order to properly model the events.

The ATLAS simulation itself is divided into three main steps. First is the aforementioned generation of a Monte Carlo event together with all the decays and hadronization. The second step is the simulation of the interaction with the detector. Finally, the digitization simulates the read-out of the detector and recording of the event, performed individually for each detector[79].

The nominal ATLAS simulation provides a precise description but has large computing requirements. This is mainly because of the generation of particle showers in the calorimeters. In order to reduce the usage of resources for analysis where a high precision is not required, an alternative *fast* algorithm is used. It is called ATLFAST (AFII)[81] and it uses a parametrization of the hadron showers instead of simulating them fully in order to reduce the computing power needed.

CHAPTER 5

Particle reconstruction and identification

In order to perform an analysis of the data recorded by the ATLAS detector, the raw information from the detector has to be interpreted to correspond to real physical objects, like muons and electrons. Their reconstruction consists of several steps combining information from the sub-detectors. Furthermore, other objects produced either in the hard scattering or coming from pile up can overlap with the reconstructed object, complicating the reconstruction. After the reconstruction, particles have to be properly selected, since some objects can have a similar response in the detector, making their identification more difficult. Finally, differences in detector response between the data and the Monte Carlo have to be corrected.

The chapter starts with a short overview of the particle signature in section 5.1, followed by a description of the track and vertex reconstruction in section 5.2. Then, section 5.3 presents the reconstruction and identification of the muons and electrons, the only leptons directly detectable by the ATLAS. Then, reconstruction of jets is discussed in section 5.4, with a large focus put on the B hadron jets and their identification through b-tagging in section 5.5. Section 5.6 gives a short description of the reconstruction of tau leptons, which only play a minor role in the measurement. Possible overlap between the reconstructed objects and how it is resolved is summarized in section 5.7.

5.1 Overview

Different particles leave diverse signatures in the detector. This is illustrated for a set of important particles in figure 5.1. As mentioned in section 4.2.2, trajectories of charged particles, also called simply *tracks*, are reconstructed from the hits in the Inner Detector. If the charged particle is a muon, it usually transverses the whole detector and leaves an additional signature in the Muon Spectometer. On the other hand electrons and photons are both stopped by the EM calorimeter; they can be distinguished from each other by trying to find a track in the ID, which would be left only by an electron. Finally, gluons and quarks produce showers of collimated particles called jets, which are detected by the hadronic calorimeter.

For electrons and muons, it is necessary to distinguish between *prompt* particles, coming directly from the interaction point or from decays of short-living particles: Z, W bosons, top quarks and tau leptons, and leptons produced in subsequent decays of long-living particles or, especially in case of electrons, produced for example in photon conversions.

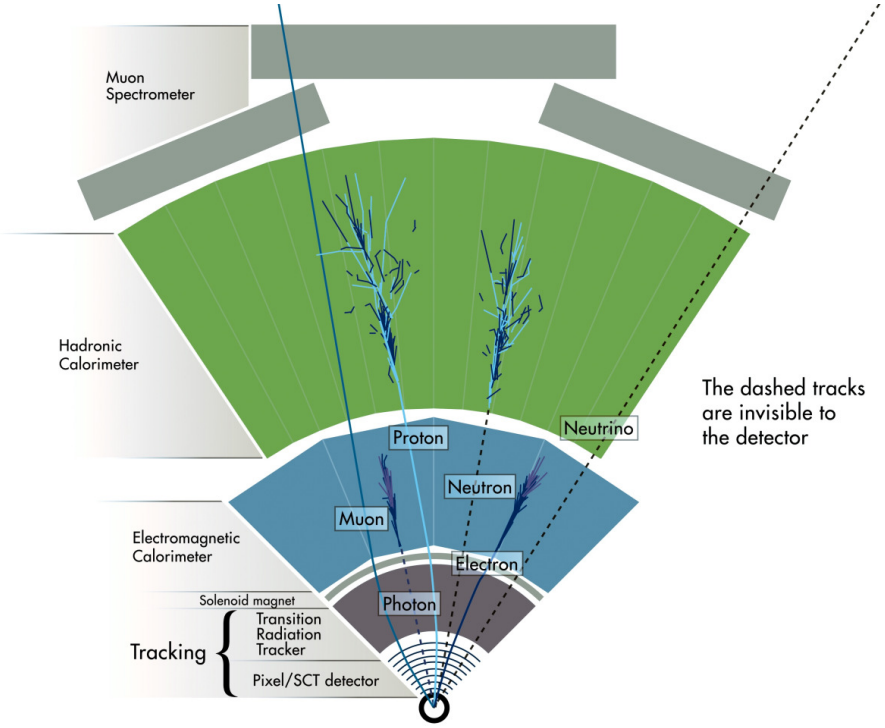


Figure 5.1: Sketch of the response of the ATLAS detector to various particles. Taken from[82], original from[83].

There are also several *derived* objects, which are constructed from the basic object like tracks or leptons. These are for example b-jets, where their identification takes advantage of the tracking information to distinguish between jets coming from a B hadron decay and other particles. Similar approach is taken for the tau leptons when decaying into hadrons. Finally, the E_T^{miss}, described previously in section 4.2.1, relies on a summation of the $\vec{p_{\mathrm{T}}}$ of all objects in the event.

5.2 Tracks and vertices

One of the most fundamental objects are tracks of charged particles reconstructed in the Inner Detector. Their trajectory, bent by the magnetic field can be used to derive their momentum. Tracks reconstructed in the ID have a good resolution, which often allows reconstruction of the vertex they originate from. That can be either a vertex of the primary interaction of protons, of a subsequent decay or of an interaction with the detector material.

5.2.1 Tracks of charged particles

The whole process of the track reconstruction is called *tracking*. The specific algorithm used for tracking in the ATLAS is called the ATLAS New Tracking (NEWT)[84]. Its basic unit is a SpacePoint, a 3D information about the place where the particle hit the detector.

All the sensors of the pixel and SCT, where the deposited energy surpasses certain predefined thresholds, are grouped together into clusters. If clusters pass certain quality criteria, they are further used in the reconstruction of particles. In case of the pixel detector, a cluster already provides a 3D information about a transition of a particle. However, the

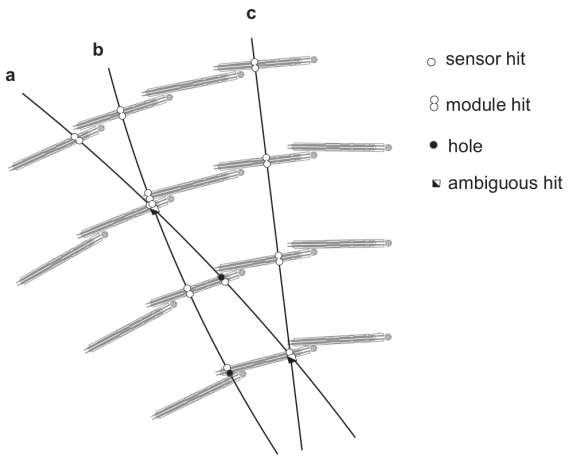


Figure 5.2: Sketch showing the basic unit of the track reconstruction in the context of the SCT subdetector[83]. Hits can either be a sensor hit or a module hit, where the latter corresponds to the case where both sensors on an SCT module were hit. If the hit corresponds to more than one track, it is called an *ambiguous* hit.

SCT only specifies a 2D position and stereo information of two modules glued together is required to create a SpacePoint.

The cluster of deposited energy is also colloquially called a *hit* (as in the place where the particle hit the detector). A point where a hit is expected from the trajectory of the particle but none is found is called a *hole*, which is an important track property when judging the quality of the reconstruction. An illustration of these concepts can be found in figure 5.2.

The tracks themselves are constructed using an *inside-out* algorithm[84], which starts by creating track candidates in the innermost layers of the detector and then extrapolates them outwards. In the first stage, a track seed is created out of sets of three SpacePoints in the pixel and SCT detectors. Three measurements offer a compromise between maximizing the number of track candidates while still allowing a first rough estimation of the track momentum and impact parameters. These properties can be used to increase purity of the seed sample[85].

In the next step, track candidates are constructed using a Kalman filter[86], extrapolating the tracks and incorporating more SpacePoints from the detector. Good understanding of the magnetic field in the Inner Detector is necessary in this stage to correctly predict the shape of the trajectory. Multiple track candidates can be constructed from a seed when more compatible SpacePoints are available.

The tracks are then assigned a score based on the number of hits and holes in the path of the track or on its χ^2. A large score is also assigned to tracks with a higher momentum, since low p_{T} tracks have a larger probability of incorrectly assigned SpacePoint. The track score is used to solve an ambiguity between overlapping tracks, a problem especially in high multiplicity events. This is done by assigning shared hits to tracks with a higher score. Tracks loosing one or more hits are then refitted. Additional selection criteria are then imposed on the tracks, which include e.g. number of hits and holes, $p_{\mathrm{T}} > 400$ MeV and $|\eta| < 2.5$ or requirements on the impact parameters. The whole procedure is described in more detail in reference [85].

The final step of the tracking is an extension into the TRT detector after which a track fit is carried out, giving the final high-resolution version of the track. Performing the track fit in the last step minimizes the CPU requirements of the track reconstructions.

5.2.2 Vertices

An important step when building an event is the reconstruction of vertices corresponding to the primary proton-proton interactions in the bunch crossing. They can be distinguished from vertices from decays of particles, such as B hadrons, or from interactions with the material where additional particles are produced.

The construction of vertices relies on the tracks already described in the previous section, done through an iterative method[87]:

1. Seed a position of a vertex from available tracks, where two and more tracks can form a seed.

2. A vertex is fitted from tracks in the vicinity of the seed through an iterative method, where less compatible tracks are removed from the fit.

3. After the vertex is found, the tracks associated with the vertex are removed from the event and the whole procedure is repeated to find the next vertex.

The vertex with the highest sum of the squared transverse momentum ($\sum_{trk}(p_{\mathrm{T}}^{trk})^2$) is called the primary vertex (PV) and corresponds to the hard scattering. In general, each physics analysis requires the presence of a PV in each event.

5.3 Leptons

Only two leptons can be detected by the ATLAS directly: electrons and muons. Since they are both charged, they leave a track in the ID. In addition, electrons rely on the EM calorimeter while the muon mainly relies on the muon spectrometer. The difficulty of their treatment does not come only from a proper reconstruction of the leptons, but also from the distinction between prompt leptons and background, composed of decays of long-living particle and of misidentified particles (e.g. charged pions).

5.3.1 Electrons

Electrons in the ATLAS detector are identified as tracks pointing to an electromagnetic cluster. They are prone to bremsstrahlung when passing the detector volume, producing an EM shower in the calorimeter as described in section 4.2.3. Because of this, multiple tracks, originating from the same primary electron, can be associated with a single energy deposition in the calorimeter. The particle showers produced this way are usually collimated.

The path of an electron through the relevant sub-detectors is depicted in figure 5.3 and a detailed description of the electron reconstruction and identification is in reference [88].

Reconstruction

The reconstruction of an electron starts in the EM calorimeter with a formation of clusters from the electromagnetic shower, taking advantage of the high granularity of the calorimeter cells ($\Delta\eta \times \Delta\phi$ =0.025×0.025). A sliding window of a size 3×5 cells is used to create *seed clusters*, if the summed energy of the cluster exceeds 2.5 GeV. In case of two overlapping seeds, only the one with a higher energy is kept if the difference in energy is higher than 10%.

The second component of the electron reconstruction are ID tracks, previously described in section 5.2.1. However, by default the tracking is performed under a pion hypothesis, since they are the most abundant particles in the detector. In contrast to pions, electrons

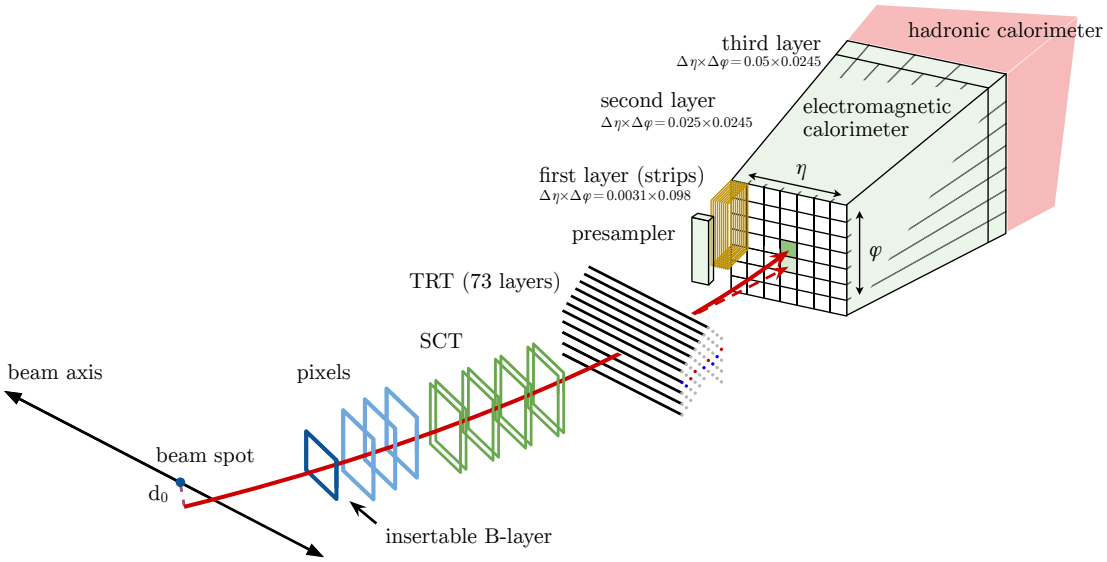


Figure 5.3: Graphic showing the path of an electron through the Inner Detector and the electromagnetic calorimeter (red solid line) and emission of an additional photon due to interaction with the detector (red dashed lines)[88].

have a larger interaction rate with the material and undergo bremsstrahlung, producing additional photons along their path. This complicates the reconstruction of their tracks.

For this reason, tracks with at least four silicon hits, which would normally not pass the track quality requirement, are refitted if they are matched to an EM cluster: the pseudorapidity is required to be $|\eta_{\text{cluster}} - \eta_{\text{track}}| < 0.05$, while the azimuthal selection depends on the charge q of the particle $-0.10 < -q \cdot (\phi_{\text{cluster}} - \phi_{\text{track}}) < 0.05$. The asymmetric selection helps to account for effects of tracks being bent more by the magnetic field due to energy loss by bremsstrahlung.

Matched tracks are refitted using the Gaussian-sum filter (GSF), which better accounts for energy losses in the material. The curvature of the track trajectory determines the momentum of the electron, and the direction of the track defines its charge.

In case several tracks are being associated with a single seed cluster, for example due to an interaction with the material, a primary track is selected by an algorithm which takes into account track-quality criteria, distance to the cluster barycenter or association with a vertex from a photon conversion.

After associating a primary track with a seed cluster, a *reconstructed cluster* is formed with an extended window of size 3×7 or 5×5 for the barrel or the end-caps respectively. The calibrated energy from the reconstructed cluster then gives the energy of the electron candidate.

The reconstruction of electrons is limited to $|\eta| < 2.47$. This is partially done to reduce noise, which would come from cells from $|\eta| > 2.5$ with a coarser granularity[89], but also due to a bad description of material in the ID, which leads to bad simulation of the electron response (this especially affects electrons because they are more prone to interaction with material).

Furthermore, the overlap region between the barrel and the end-caps ($1.37 < |\eta| < 1.52$) is excluded in most physics analyses due to extra material and the difficulty of matching and calibrating energy from cells of different geometries and orientation[90].

Identification

The main purpose of the identification step is to further distinguish the prompt electrons from the background. To achieve this, a likelihood (LH) discriminant[1] is constructed, taking advantage of both the tracking and calorimeter information. Numerous variables are used, including the shape and the size of the EM shower, the number of hits in the various layers of the Inner Detector or the matching of tracks to the cluster. Information from the TRT is used as well, since the transition radiation can be used to differentiate between electrons and pions.

The full list of variables used in the likelihood discriminant and the technical implementation can be found in reference [88]. Several working points, defined through selection on the LH discriminant, are then used to create samples of reconstructed electrons. Tighter selection means less background but also a lower selection efficiency for electrons. Whether it is more desirable to reject more background or to select more electrons highly depends on the analysis strategy.

In the Run-2 of the ATLAS data collecting, three main WPs are defined, based on the identification efficiency of an electron at $E_t = 40$ GeV. Starting with the highest efficiency there is the *Loose* (93%), the *Medium* (88%) and the *Tight* (80%) working point. It is important to note that the efficiency depends slightly on $|\eta|$ and strongly on the E_t of the electron.

Isolation

The next step in the treatment of electrons is the isolation. Its main purpose is to further discriminate the prompt leptons from various backgrounds (non-prompt electrons and misidentification) which typically have a larger activity in the proximity of the electron. For example, in case of an electron from a B hadron decay there will be either a jet or at least additional tracks in the vicinity. Hence, an isolated electron is more likely to be a prompt lepton. Unlike the identification, which takes into account properties of the reconstructed electron, the isolation relies on its relation to other objects. In practice, this means looking at a certain radius ΔR around the electron and summing the transverse energy (in case of the calorimeter) or the momentum (in case of tracks), excluding the energy/momentum of the electron itself.

In case of the former, there is some ambiguity to what energy is deposited by the electron and which by other particles. The subtraction of an electron energy is therefore simplified by removing cells within a rectangle of size $\Delta\eta \times \Delta\phi = 0.125 \times 0.175$. In addition, the average energy leaking outside this area is estimated from the Monte Carlo and subtracted as well.

The energy deposited by the pile-up has to be calculated as well. This is done by measuring the overall density of energy deposition in the detector in each event. The area considered in the isolation is defined by $\Delta R = 0.2$ from the electron candidate. Lower values are not practically accessible due to the granularity of the EM calorimeter.

The track-based isolation only considers tracks with $p_{\mathrm{T}}^{e} > 5$ GeV, passing additional requirements on the longitudinal IP z_0 ($|z_0 \sin\theta| < 3$ mm), suppressing tracks from the pile-up significantly. Then, particle tracks within $\Delta\eta \times \Delta\phi = 0.05 \times 0.1$ around the electron candidate are removed under the assumption that they are mostly coming from the electron and its radiation. The selected area used in the isolation selection can be more flexible

[1] The likelihood is described later in section 7.1.1 in the context of maximum likelihood fits. For now, its sufficient to note that a likelihood is constructed out of probability functions to quantify the compatibility between a prediction and the observed data. Likelihood is, however, not a probability distribution in itself.

compared to the calorimeter due to the higher granularity of the ID. It is defined through a ΔR radius, determined by the following formula:

$$\Delta R_{iso} = \min\left(\frac{10\text{GeV}}{p_{\text{T}}^{e}[\text{GeV}]}, 0.2\right).$$

There are several isolation working points used in the ATLAS, based on the isolation efficiency. The whole set of WPs can be found in [88], but in the context of the $t\bar{t}H(b\bar{b})$ analysis only two are important. The *Loose* isolation, which simply requires 99% isolation efficiency[2] for both the tracking and calorimeter component, and the *Gradient* isolation, which has a p_{T} dependent requirement on the efficiency $\epsilon_{iso} = 0.057 \times p_{\text{T}}[GeV] + 95.7\%$ up to 99% at p_{T} =60 GeV. Above 60 GeV the efficiency requirement is kept constant at 99%.

Calibration

The simulation of the detector used in the Monte Carlo is not perfect and the differences in electron response between the data and the MC have to be corrected. This is done by studying Z and J/Ψ decays into a pair of electrons. By focusing on a narrow peak in the invariant mass of the two leptons, a pure sample of electrons can be obtained. One can then can be selected with a tighter requirement and used to select, or in other words *tag*, the event. The second electron with a looser selection is then used to *probe* the properties of the electron reconstruction and identification. The method is then called *tag&probe* method. The differences in the efficiency of the reconstruction, identification and isolation between the data and the Monte Carlo are studied and corrected using an event weight[91].

5.3.2 Muons

Muons have a low interaction rate with the material and usually travel through, and even leave, the whole volume of the detector. As such, they are the only particle detected by the Muon Spectometer[3]. The MS is used both to identify the muons and to improve the precision of the track reconstruction in comparison to using the ID. The whole procedure of muon reconstruction and selection is described in more detail in reference [92].

Reconstruction

The reconstruction of muons relies mainly on combining an ID track with information from the Muon Spectometer. Several types of muons are defined. The ID track can be combined either with a full track from the Muon Spectometer (*Combined muon*), or a segment of a track (*Segment-tagged muon*). In addition, a track is identified as a *calorimeter-tagged muon* if it can be connected with an energy deposition in the calorimeter system which matches a minimum ionizing particle (MIP)[21]. The final type is the *extrapolated muon*, which matches an MS track to the interaction point. It is mainly used in the forward $2.5 < |\eta| < 2.7$ region not covered by the ID. The $t\bar{t}H(b\bar{b})$ analysis uses only the Combined (CB) muons.

Tracks in the Muon Spectometer are reconstructed in a similar way as tracks in the ID. The main difference is that instead of starting in the inner layers and performing an inside-out extrapolation, the muon tracking starts in the middle layers of the MS and is extrapolated both inwards and outwards. A global χ^2 fit is performed on the associated hits, where hits with a large contribution to the χ^2 are iteratively removed.

[2]Here, isolation efficiency means selection efficiency of a prompt lepton coming from a Z decay.
[3]In some cases particle with high energy can reach the spectrometer, but the rate is small.

For the combined muons, independent tracks in the ID and MS are constructed. Then, the latter is extrapolated into the ID and a matching track is found. A complimentary inside-out extrapolation is subsequently used, though it only contributes to a minority of reconstructed muons.

Identification

The identification step is used to suppress the background from pion and kaon decays, containing a non-prompt muon, and to ensure a good momentum resolution. Decays producing the non-prompt muon are identified by a kink in the muon trajectory and can be excluded by a tighter requirements on the fit quality. The momentum defined individually from the ID and MS track segments is also compared, where a decay will affect the p_{T} in the Inner Detector.

Several working points are defined to offer a flexibility of the muon identification efficiency on one side and the background rejection on the other. First, there are *Loose, Medium* and *Tight* WPs, which have a muon identification efficiency of approximately 98%, 96% and 92% for muons with $20 < p_{\mathrm{T}} < 100$ GeV[92]. In addition, a *High-*p_{T} identification is defined for muons above 100 GeV, where the emphasis is on a good momentum resolution in exchange for a lower identification efficiency (80%).

Isolation

To further distinguish prompt muons from those originating from e.g. semileptonic decays, additional requirements on a low activity of particles around the lepton is used. Similarly to the electrons, these isolation requirements are divided between the tracks and calorimeter clusters.

The calorimeter activity is defined by the energy deposited in a $\Delta R = 0.2$ distance around the muon, from which the energy of the muon and contribution from the pile-up have to be subtracted. Since muons are minimum ionizing particle, they do not contribute as much to the calorimeter deposition (in contrast to the electrons).

The track-based isolation is formed through a more flexible ΔR definition with respect to the calorimeters, depending on the p_{T} of the muons:

$$\Delta R_{iso} = \min\left(\frac{10\mathrm{GeV}}{p_{\mathrm{T}}^{\mu}[\mathrm{GeV}]}, 0.3\right),$$

where the p_{T} of all tracks is summed (defined as p_{T}^{vc30}), excluding only the muon track.

The isolation working points are presented in detail in reference [92]. Only two are relevant to the $t\bar{t}H(b\bar{b})$ measurement, the *Gradient* WP, which requires a p_{T} dependent isolation efficiency, starting 90% at 5 GeV and going up to 99% for muons with $p_{\mathrm{T}} > 60$ GeV, and the *FixedCutTightTrackOnly*, which relies only on the track component of the activity definition and is defined by a simple inequality on the summed track p_{T}: $p_{\mathrm{T}}^{vc30}/p_{\mathrm{T}}^{\mu} <= 0.06$.

Calibration

Similarly to the electrons, the muon performance is studied in decays of Zs and J/Ψs, where a pure sample of muons can be studied using the tag&probe method. Differences in the efficiency between the data and the Monte Carlo are then corrected for all stages of the muon reconstruction and identification.

5.4 Jets

As was already established in the context of the QCD in section 2.3, color-charged particles like quarks and gluons cannot be observed directly. Instead, additional particles are produced through fragmentation and hadronization until a stable color-less final state is achieved. The resulting shower of collimated particles is called a jet.

5.4.1 Jet definition and algorithms

A jet is a practical way to measure and quantify the plethora of collimated particles produced in processes involving gluons and quarks. The reconstructed jet ideally describes the main properties of the original parton, its direction and energy.

Jets are defined through jet algorithms. More detailed description can be found for example in reference [93], here only their main properties and the primary algorithm used in ATLAS, the anti-k_{T} algorithm[94], are discussed.

The first important property of a jet algorithm is that it gives a good description on both theoretical level (particle jets) and on experimental level (reconstructed jets), allowing a direct comparison between them.

The next two important attributes have to do with an additional radiation. A good jet algorithm has to be *collinear* and *infrared* safe, meaning that additional emissions of *collinear* and *soft* particles should not affect the number and properties of reconstructed jets. The former can be a problem of jets defined around the most energetic particles in the event. Particles undergoing a collinear splitting can lose a significant part of their energy and instead of seeding a new jet can get absorbed into an another nearby jet (or no jet get reconstructed at all e.g. due to an energy threshold). The infrared safety is connected to low energy emissions in larger angles. This for example concerns nearby jets, where a soft radiation in the phase-space between them can lead to the jet algorithm merging the jets. Both concepts are illustrated in figure 5.4

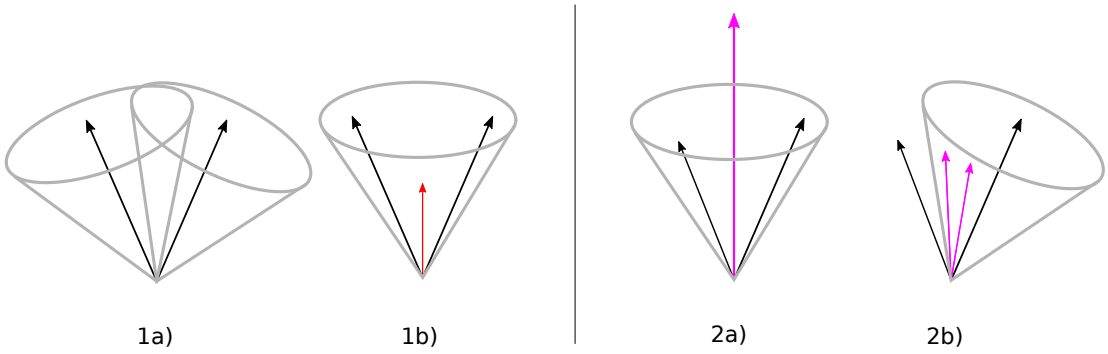


Figure 5.4: Sketch illustrating (1) an infrared and (2) a collinear radiation and its possible undesired impact on the jet reconstruction. The arrows represent particles and the grey cones the reconstructed jets. 1a) shows two individual jets which become merged in 1b) when additional soft radiation in-between them takes place. 2a) shows a jet centered around the most energetic particle, which in case of collinear splitting 2b) has much smaller energy and the second most energetic particle becomes the center of the jet, significantly shifting the direction of the jet while also excluding one of the particles.

Though numerous jet algorithms exist, only the anti-k_{T} algorithm, discussed in the next section, is used in the ATLAS. For more examples one can once again consult the reference [93].

Anti-k_{T} algorithm

The anti-k_{T} algorithm[94] clusters nearby objects until a certain condition is met. The relevant distance between two entities i, j is defined as

$$d_{ij} = \min(p_{\mathrm{T},i}^{-2}, p_{\mathrm{T},j}^{-2}) \frac{\Delta_{ij}^2}{R^2}, \tag{5.1}$$

where $p_{\mathrm{T},i}$ is the transverse momentum of the given object, $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ is simply the distance in $y - \phi$ plane, and R is a parameter of the algorithm. In addition, the distance to the beam, defined simply as $d_{iB} = p_{\mathrm{T},i}^{-2}$, serves as a cut-off value.

These distances are computed for each object in the event and the smallest one is found. If it is the d_{iB}, the entity i is removed from the event and is defined as a jet. Otherwise, if it is the d_{ij}, the two objects are merged. The algorithm continues until there are either no other particles in the event or some cut-off value on the distance is reached.

The anti-k_{T} algorithm is collinear and infrared safe. Collinear particles have by definition a small distance to the other particle and will be merged among the first, creating the same object as if the splitting never took place. On the other hand, soft particles are just assigned to the nearest hard jet (as long as they are within R distance in the $y - \phi$ plane) and do not affect the merging or seeding of jets.

5.4.2 Energy clusters

In the context of the ATLAS reconstruction, the objects used by the anti-k_{T} algorithm are topological clusters (or topo-clusters), constructed from individual hadronic calorimeter cells. Since each cell produces at least some signal due to the detector noise, they are first assigned a signal significance $\xi_{cell} = E_{cell}/\sigma_{E,cell}^{noise}$, where E_{cell} is simply the energy measured in a given cell and $\sigma_{E,cell}^{noise}$ is the average noise.

The topo-clusters are then seeded in each cell with $\xi_{cell} \geq 4$. All neighbouring cells are then added to the cluster. The procedure is then repeated for all cells included in the cluster with $\xi_{cell} \geq 2$ until there are no more cells left to add. If a cell could be assigned to two clusters, the clusters are merged. The cell energies are measured at EM scale, which assumes that the energy was deposited by an electro-magnetically interacting particles.

5.4.3 Small-R jets

Small-R jets, usually referred to simply as jets, are reconstructed from the topo-clusters using the previously mentioned anti-k_{T} algorithm implemented in the FastJet 2.4.3 software package[95] with the value of the radius parameter set to $R = 0.4$.

Calibration

Reconstructed jets are subject to numerous corrections to better match the theoretical prediction and to improve the agreement between the data and the Monte Carlo[96]. In the first step, the jet direction is changed to point to the primary vertex, affecting its direction but keeping the same energy.

The measured hard-scatter event is accompanied by particles from pile-up, which may overlap with the jet and increase its energy. To estimate their contribution, the median p_{T} density of jets[4] in the detector volume within $|\eta| < 2$ is measured. Subsequently p_{T} is subtracted from the jet based on its area in the $y - \phi$ and the average density. This

[4] p_{T} density is defined as p_{T} of the jet over its area.

procedure is accompanied by a residual pile-up correction which is derived as a function of μ and nthe umber of primary vertices N_{vtx}.

In the next step, the four-vector of the jet is corrected to match the particle level jet using an MC-driven calibration. This correction is called jet energy scale (JES) and corrects the energy and direction of the jet to match their particle level values.

Jet properties are sensitive to the type of the initial particle. For example, a jet initiated by a gluon has on average a larger number of softer particles, leading to a lower calorimeter response. The impact of this dependency is fixed through a global sequential calibration (GSC), which identifies variables sensitive to the initial state and corrects them sequentially.

Finally, differences between the data and the Monte Carlo due to imperfect detector modeling have to be resolved. This is done using an *in situ* calibration method, which takes advantage of the momentum conservation: the p_{T} of a jet has to be balanced with other objects in the event so the total transverse momentum is zero. A well calibrated object is used as a reference. This can be for example a Z boson (identified as a pair of muons or electrons with an invariant mass Z) or a photon. Forward jets (with $|\eta| > 2.5$) are also calibrated in dijet events where the other jet is central. There are two categories of *in situ* corrections[96, 97]: *in situ* jet energy scale and *in situ* jet energy resolution (JER) calibration. The first further corrects the energy in the data (in the MC it is already calibrated to match the particle level) and the second smears jets in the Monte Carlo so their resolution better matches one of the jets in the data.

5.4.4 Jet Vertex Tagger

The goal of the jet vertex tagger (JVT)[98] is to suppress jets originating from the pile-up interactions. It takes an advantage of tracks associated with the jet to create a discriminant which determines whether a given jet is more likely to come from the primary vertex or a pile-up interaction.

It uses two main variables. One is connected to the fraction of the transverse momentum carried by tracks associated with the PV compared to the transverse momentum of all tracks. In case of jets from the pile-up this fraction should approach zero. The second variable describes the fraction of p_{T} of the jet carried by the tracks associated with the primary vertex.

5.4.5 Additional jet types

The jets described so far are part of a group called *EMTopo* jets (EM for the EM calibration and Topo since the topo-clusters are the basic units used in the algorithm). In addition, there are *particle flow* (*PFlow*) jets[99].

The basic object of the PFlow jet is not a cluster. Instead, it combines tracks with clusters in the calorimeter to create more complex objects to better distinguish between the charged and neutral particles. The matching to the primary vertex also allows to subtract part of the pile-up already before the jet reconstruction, compared to the EMTopo jets where pile-up subtraction and suppression takes place after the reconstruction. PFlow jets are described in detail in reference [99].

5.5 B-tagging

There are four b quarks in the final state of the $t\bar{t}H(b\bar{b})$ process. However, jets described in the previous section do not differentiate the particle which produced the jet, making the selection of the process of interest practically impossible. For this reason, a flavor tagging

procedure tries to determine the primary particle of the jet. Its main purpose is to identify B hadron jets (or b-jets) and for that reason it is called *b-tagging*. It also differentiates between jets originating from a D hadron (or c-jets) and the remaining jets, simply called *light jets*. More details about the ATLAS b-tagging can be found in references [100, 101].

B hadrons have one significant property, which allows their differentiation from other jets. They have a relatively long lifetime (1.5ps), large enough to travel few millimeters from the primary vertex, but usually not long enough to actually reach the detector. This unique signature manifests itself through a number of properties in the reconstructed event.

Primarily, the ID tracks associated with the b-jet are usually further from the primary vertex, since the secondary decay vertex which is their origin is few mm away from it. This results in larger impact parameters d_0 and z_0. Furthermore, the tracks will usually intersect the jet axis in the transverse plane on one side of the primary vertex in direction of the jet. Finally, it is often possible to reconstruct the secondary vertex of the B hadron decay.

Tracks are associated with a jet if they can be found within a certain ΔR distance from the jet axis. This requirement is dependent on the jet p_{T}, since boosted jets usually have more collimated particles, resulting in a narrower jet. The impact parameter used for b-tagging is then signed based on the point where they cross the jet axis with respect to the primary vertex: it is positive if it is in direction of the jet and negative otherwise.

All the main properties of a B hadron jet and its associated tracks are displayed in the figure 5.5.

5.5.1 B-tagging algorithms

Algorithms responsible for the b-jet identification are divided into two parts. The low level algorithms are trying to either distinguish between b/c/light jets based on impact parameters of their tracks, or they attempt to reconstruct the jet properties, like the secondary vertex or the whole decay chain. The high level algorithms then combine the low-level outputs to create a high-performance b-tagging discriminant.

Low-level algorithms

Two low-level algorithms, working based on the distance between tracks and vertices, are the IP2D and IP3D algorithms[100]. The former constructs a likelihood based on the signed transverse impact parameters of tracks associated with the jet. The ratio of logarithmic likelihoods (LLR) between each of the three jet flavors then provides the first low-level discriminant. The IP3D then further incorporates the transverse impact parameter z_0 and its correlation to d_0.

The SV1 algorithm[102] aims to reconstruct the secondary vertices within the jet. They are seeded by combining pairs of tracks, excluding tracks and vertices compatible with long-living particles (e.g. K_s), or those from photon conversions or material interactions. All nearby tracks are fitted to form a vertex candidate, iteratively removing tracks with a high χ^2 until a single stable vertex is created. Several techniques are implemented to reduce the impact of pile-up tracks, for example limiting the number of tracks used in the reconstruction to 25 tracks with the highest p_{T}. Properties of the secondary vertex, like for example its invariant mass or number of associated tracks, are then used in the high-level algorithms.

Finally, the topological multi-vertex algorithm JetFitter[103] aims to reconstruct the decay chain of B and D hadrons, the former giving two displaced vertices (one for decay of the b and the second for the subsequent decay of the D hadron), while the latter

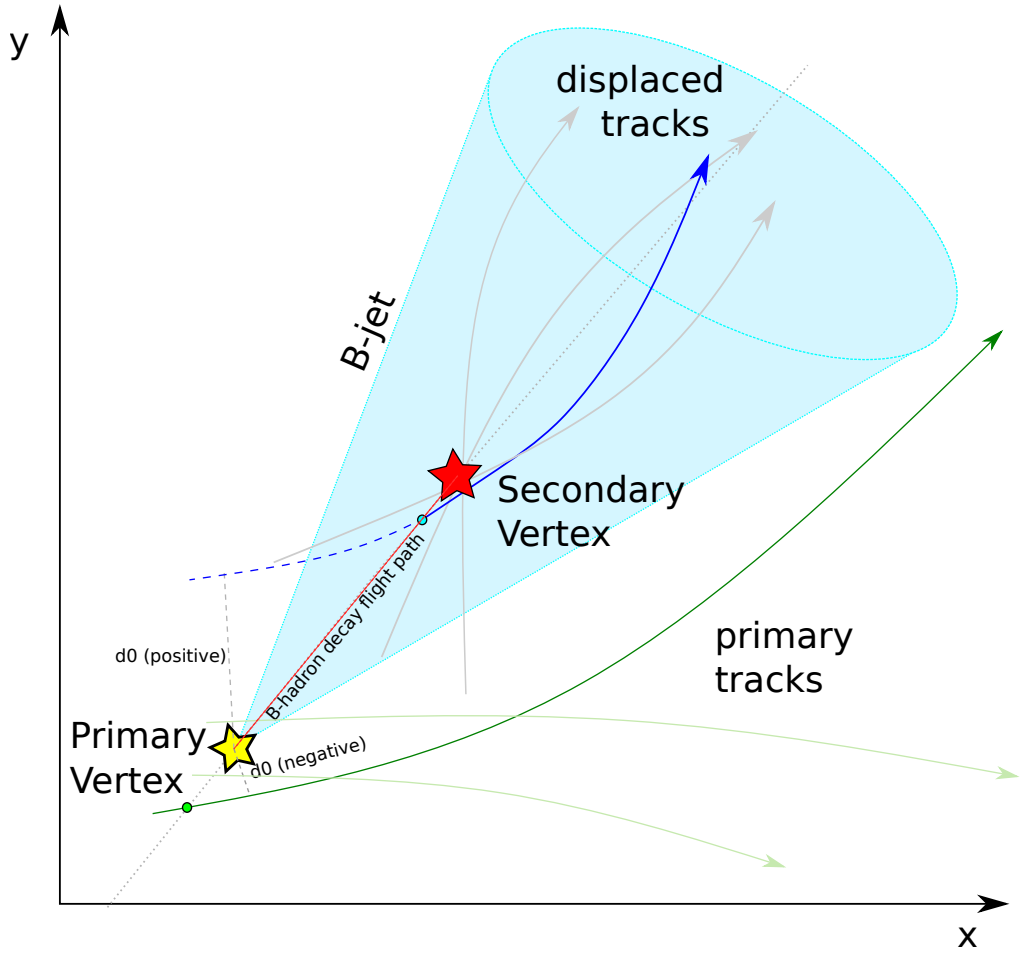


Figure 5.5: Sketch displaying the main properties of a jet originating from a decay of a B hadron, shown in the transverse plane with respect to the beam-pipe. The main properties are the presence of the secondary vertex due to relatively long flight path and larger positive impact parameters of the associated tracks. The impact parameter is positive when the track crosses the jet axis on the side of the primary vertex in direction of the jet (when projected in the transverse plane).

b-jet efficiency [%]	c-jet rejection	light jet rejection
60	34	1538
70	12	381
77	6	134
85	3.1	33

Table 5.1: b-jet efficiency at the four working points of the b-tagging and the corresponding rejection of background from c jets and light jets[101].

only one. Properties of the resulting vertices serve as additional inputs of the high-level discriminators.

High-level algorithms

Two high-level algorithms are currently implemented in the ATLAS experiment, both based on multi-variate algorithms. First, there is the older *MV2*, based on Boosted Decission Tree (BDT) (described in appendix A). It assumes events with b-jets as signal while the background is composed of a certain fraction of c and light jets. In the case of the $t\bar{t}H(b\bar{b})$ analysis, specific *MV2c10* version is used, which assumes 7% of the background coming from c-jets[101], the rest coming from light jets.

The second high-level algorithm, called *DL1r*, is based on Deep Neural Network (DNN) (for more information see reference [104]) based on an older algorithm *DL1*[105] with an improved performance compared to the MV2 algorithm[106]. It has a higher discriminating power, but because its calibration was not available at the time of the analysis presented in this thesis, it is not considered. Both algorithms provide a single output, a b-tagging discriminant. Higher values of the discriminant represent a higher chance of a given jet being a product of a B hadron decay. The remainder of this section will focus on the performance of the MV2c10 algorithm.

5.5.2 Working points

Similarly to the identification of electrons and muons, several working points are constructed for the b-jet selection. They are based on the efficiency of correctly ascribing the b-jet label with four working point: 85%,77%,70% and 60%. The rejection of the c and light jet background can be found in table 5.1. It clearly demonstrates the trade-off between a lower efficiency and a higher rejection rate. This is especially important for analyses incorporating multiple b-jets, which is the case of the $t\bar{t}H(b\bar{b})$ analysis with four b-jets in the final state. Selecting four b-jets at 60% WP will yield a relatively clean b-jets sample but the efficiency of the selection will be staggeringly low: $\epsilon_{sel} = 0.6^4 \approxeq 13\%$!

5.5.3 Calibration

The response of the b-tagging algorithm in the Monte Carlo is calibrated so it agrees with the data. This is done separately in $t\bar{t}$ events for b-jets [107] and c-jets [108], where c-jets are investigated in a final state where one of the W decays into cs, and for light jets in di-jet events[109]. These calibrations are generally done with a specific Monte Carlo, however, the response can vary greatly between generators. For this reason additional MC-MC corrections are derived[110].

5.6 Tau lepton

5.6.1 Leptonic and hadronic decays of tau lepton and top quark

In order to discuss tau leptons in this chapter and top decays later on, it is important to define the concept of *leptonic* and *hadronic* decays in context of an ATLAS measurements. It is not based on the direct products of the associated W decay, but rather on the resulting signature in the detector. If the W decays into e, μ (+their neutrinos), the decay is considered to be leptonic, while for quarks it is hadronic.

When a W from a top quark decays into a tau lepton (and its neutrino), the situation is slightly more complicated. Tau leptons decay too quickly to be detected by the detector. Therefore, when a tau is concerned, the W decay is classified based on the classification of the subsequent tau decay.

5.6.2 Leptonic tau lepton

Tau lepton decaying into electrons and muons cannot be directly distinguished from other prompt leptons, since the only other particle produced in their decay, the neutrino, is not detected by the detector. For this reason, there is normally no special treatment for leptonic tau leptons.

5.6.3 Hadronic tau lepton

The hadronic final state represents 65% of the decay modes of the tau leptons and typically has one or three charged pions in the final state. For the former, an electron misidentification is the main background, since it similarly has only a single track in the ID. For the latter, the background is dominated by gluon and quark jets.

Tau leptons are seeded from small-R jets. The seed with the largest fraction of momentum carried by tracks with $p_{\mathrm{T}} > 1$ GeV and within $\Delta R < 0.2$ from the center of the jet is considered to be a tau lepton candidate. All candidates are then calibrated for a presence of a pile-up and to match the true value of energy from the Monte Carlo.

A BDT is then used to create an identification discriminant to reject background from quark and gluon jets, based mainly on properties of the associated tracks. Additional electron discrimination is applied on tau lepton candidates with a single associated track.

The tau leptons are further calibrated using $Z \to \tau\tau$ events to mitigate differences between the data and the Monte Carlo simulation. More information on the tau lepton reconstruction and calibration can be found in reference [111].

5.7 Overlap removal

A single detector signature can be associated with multiple objects. To avoid double counting of the signatures, an overlap removal procedure is applied for each event. First, jets within $\Delta R_y = 0.2$ distance from an electron are removed, reducing the number of jets reconstructed from electron energy depositions in the calorimeter. Then, electrons and muons within $\Delta R_y = 0.4$ from any remaining jet are also removed. This reduces the background of non-prompt electrons from heavy flavor decays. For muons there is an exception in case the associated jet has less than three tracks where the jet is removed instead. Such jet often comes from energy deposition of a high p_{T} muon. Finally, a hadronic tau lepton candidate is removed if it is within $\Delta R_y = 0.2$ from an electron or muon.

CHAPTER 6

Search for $t\bar{t}H(b\bar{b})$ in the single lepton channel

The $H \to b\hat{b}$ decay channel dominates among the $t\bar{t}H$ final states, accounting for almost 60% of the production. Its measurement is difficult due to the presence of an irreducible background from a top quark pair production with two additional b quarks in the final state coming from a splitting of a gluon emission ($t\bar{t}b\bar{b}$). This process is difficult to model due to a high number of jets in the final state and presence of heavy particles with significantly different masses.

The $t\bar{t}H(b\bar{b})$ measurement is divided into multiple channels based on the number of leptons in the final state. In this chapter the general properties of the $t\bar{t}H(b\bar{b})$ analysis in a single lepton final state are introduced. Chapter 7 then discusses the statistical analysis of this channel, which consists mainly of the signal extraction through a profile likelihood fit. Properties of the other channels and specifically of the dilepton final state will be discussed in chapter 8 in the context of a combined measurement.

This chapter starts with a short summary of the previous ATLAS analysis of the $t\bar{t}H(b\bar{b})$ process in section 6.1. The new measurement and its prospect are presented afterwards in section 6.2. Introduction and discussion of the analyzed dataset is given in section 6.3, section 6.4 then presents the Monte Carlo modeling, with an emphasis on the dominant $t\bar{t}b\bar{b}$ background. Then, the analysis regions are defined and the main properties of the $t\bar{t}H(b\bar{b})$ final state and the background composition are discussed in section 6.5. Variables used in the fit are also introduced in section 6.6. Section 6.7 is reserved for a discussion of various techniques used to treat the input distributions and the systematic variations used in the statistical analysis. Finally, the agreement between the data and the nominal theoretical prediction is discussed in section 6.9.

6.1 Previous ATLAS $t\bar{t}H(b\bar{b})$ measurement at 13 TeV

The first ATLAS $t\bar{t}H(b\bar{b})$ analysis at 13 TeV[8] was performed on the data collected in 2015 and 2016 corresponding to an integrated luminosity of 36 fb^{-1}, studied in the single lepton and dilepton channels. The signal strength $\mu_{t\bar{t}H}$, defined as a ratio between the measured and predicted cross-section, was extracted from the data using a profile-likelihood fit, leading to the result shown in figure 6.1. The $\mu_{t\bar{t}H}$ is displayed separately for the two leptonic channels and for their combination. The result is in agreement with the Standard Model and gives an upper limit of $\mu_{t\bar{t}H} < 1.2$ in absence of the $t\bar{t}H$ signal at 95% confidence level[8].

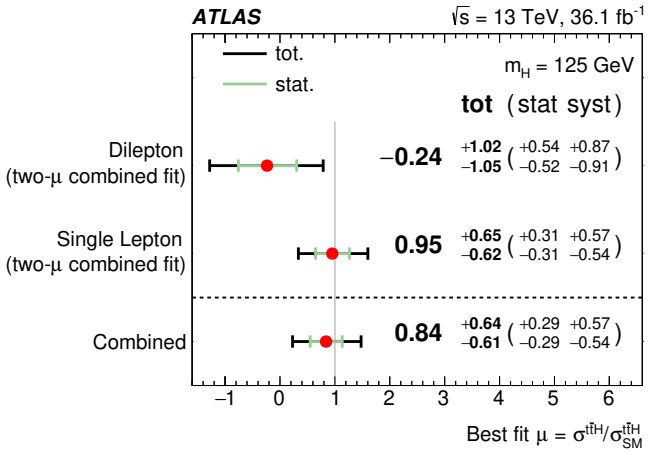


Figure 6.1: Comparison of signal strength μ, designated as $\mu_{t\bar{t}H}$ in the text, measured in the two channels separately (but with correlated systematic uncertainties) and their combination in the 2018 $t\bar{t}H(b\bar{b})$ measurement[8].

The two largest sources of systematic uncertainties limiting the sensitivity of the measurement were the modeling of the $t\bar{t}$ with additional heavy flavor jets, discussed later in section 6.4.3, and the limited size of the Monte Carlo samples, which are used to estimate the background and its systematic uncertainties. The contribution of different systematic sources to the uncertainty of $\mu_{t\bar{t}H}$ can be seen in table 6.1.

Uncertainty source	$\Delta\mu$	
$t\bar{t}+\geq 1b$ modeling	+0.46	−0.46
Background-model stat. unc.	+0.29	−0.31
b-tagging efficiency and mis-tag rates	+0.16	−0.16
Jet energy scale and resolution	+0.14	−0.14
$t\bar{t}H$ modeling	+0.22	−0.05
$t\bar{t}+\geq 1c$ modeling	+0.09	−0.11
JVT, pileup modeling	+0.03	−0.05
Other background modeling	+0.08	−0.08
$t\bar{t}+$ light modeling	+0.06	−0.03
Luminosity	+0.03	−0.02
Light lepton (e,μ) id., isolation, trigger	+0.03	−0.04
Total systematic uncertainty	+0.57	−0.54
$t\bar{t}+\geq 1b$ normalization	+0.09	−0.10
$t\bar{t}+\geq 1c$ normalization	+0.02	−0.03
Intrinsic statistical uncertainty	+0.21	−0.20
Total statistical uncertainty	+0.29	−0.29
Total uncertainty	+0.64	−0.61

Table 6.1: Summary of main sources of systematic uncertainties in the 2018 $t\bar{t}H(b\bar{b})$ measurement [8].

6.2 Full Run-2 analysis of $t\bar{t}H(b\bar{b})$

At the end of the LHC Run-2 data-taking (see section 4.1.2), a 139fb^{-1}dataset of proton-proton collisions was available for an analysis at 13 TeV center-of-mass energy. For the $t\bar{t}H(b\bar{b})$ channel, the benefit from the increase of the data statistics is limited, since the analysis has large systematic uncertainties. There are, however, several areas of improvement:

- Alongside the larger data sample size, more Monte Carlo events were produced, leading to **increased statistics of the simulated samples** not only for better estimation of the nominal background, but also of its systematic uncertainties.
- Due to better understanding of the detector and of the reconstruction, **smaller instrumental uncertainties** are expected. Their improved performance implies that more effort can be committed to investigation of the background modeling.
- A **new nominal generator for the $t\bar{t}b\bar{b}$ process**, which contains the two b quarks directly in the matrix elements. This is discussed in more detail in section 6.4.3.

The aim of the Run-2 analysis is to produce a result with a full Run-2 dataset with an improved precision of the measured signal strength.

This thesis shares many aspects with a measurement presented in reference[112], which uses a modified strategy to focus on differential properties of the $t\bar{t}H$ process.

6.3 Dataset and trigger requirements

The data collected by the ATLAS experiment are selected by triggers (see section 4.2.6). In the $t\bar{t}H(b\bar{b})$ analysis, they have been specifically collected using single electron[113] and single muon[114] triggers. Each trigger is defined by a p_{T} threshold of the lepton and by identification and isolation requirements (described in chapter 5). All triggers used in the analysis are listed in table 6.2. Due to the different conditions in the first year of the Run-2 data-taking, mainly a lower instantaneous luminosity which allows for lower thresholds, the triggers are different from those used after 2015.

6.4 Modeling and Monte Carlo generators

The analysis uses numerous MC samples to construct the nominal model of the signal and the background and to estimate the systematic variations. In this section, common properties among them are discussed, before presenting specific generators and their purpose. Heavy flavor classification, which classifies jets based on the particle they originate from, is also defined. Detailed description of the main properties and principles of MC generators was already presented in the context of the event generation throughout chapter 3.

6.4.1 Common treatment of the MC samples

The pile-up (see section 4.1.1) is simulated the same way in all samples using the Pythia8 generator[115]. Parameters of the model were tuned using minimum-bias data collected during 2015[116], resulting in a tune called A3[117].

Pythia8 is also the nominal parton shower and hadronization generator for the samples used in the analysis. A specific A14 tune[118] was developed by ATLAS to better describe showers of hard scattering events. The alternative showering generator is Herwig7[40,

Year	Trigger name	Object	p_{T} [GeV]	ID	Isolation
2015	e24_lhmedium_L1EM20VH	Electron	≥24	medium	gradient
	e60_lhmedium		≥60	medium	-
	e120_lhloose		≥120	loose	-
	HLT_mu20_iloose_L1MU15	Muon	≥20	loose	-
	HLT_mu50		≥50	-	-
2016 - 2018	e26_lhtight_nod0_ivarloose	Electron	≥26	tight	loose
	e60_lhmedium_nod0		≥60	medium	-
	e140_lhloose_nod0		≥140	loose	-
	HLT_mu26_ivarmedium	Muon	≥26	medium	gradient
	HLT_mu50		≥50	-	-

Table 6.2: Single lepton triggers used to select events in the leptonic channels, showing the selected object, its p_{T} threshold, the identification and isolation working points (previously described in chapter 5), divided into two periods - 2015 and 2016-2018. The triggers are explained in more detail in references [113] and [114] for electrons and muons, respectively. An event is considered passing the trigger if it passes any of the listed triggers.

119] using the H7UE tune[40]. For both showering generators, the B and D hadron decays are simulated using EvtGen [120]. For all samples, the mass of the top quark (m_{top}) is set to 172.5 GeV.

6.4.2 Heavy flavor classification

The dominant background of the $t\bar{t}H(b\bar{b})$ analysis is $t\bar{t}b\bar{b}$: a $t\bar{t}$ process with two additional b quarks coming e.g. from an emission of a gluon which further splits into two b-quarks. When $t\bar{t}$ is generated inclusively with additional jets from the parton shower, the $t\bar{t}b\bar{b}$ component has to be subsequently filtered.

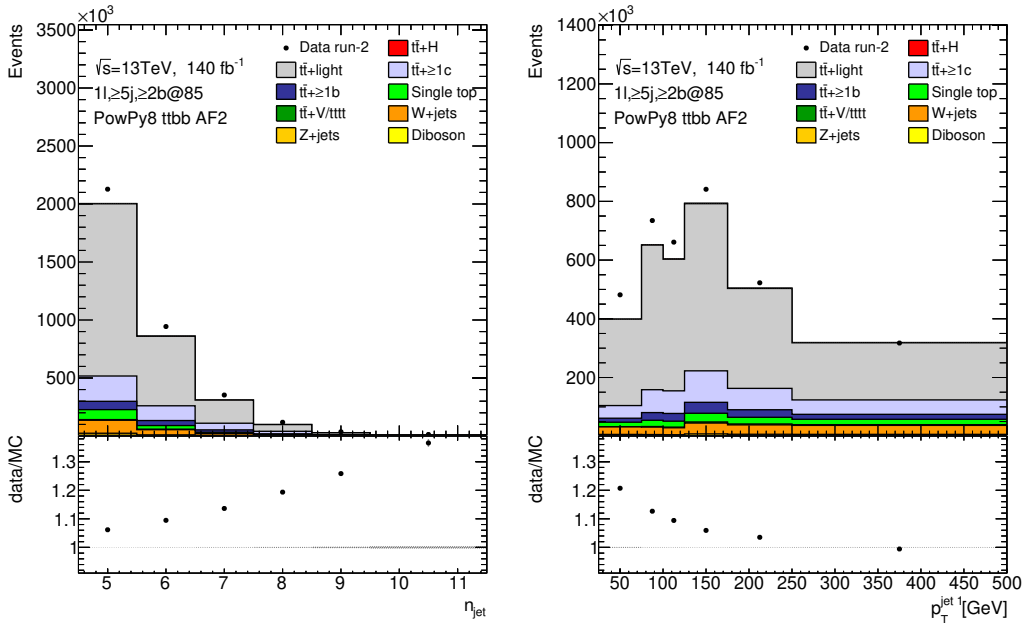
The heavy flavor classification[121] finds the parton of origin for jets in the MC samples. First, particle level[1] jets of the MC are reconstructed using an anti-k_t algorithm, described in section 5.4.1, with a radius R=0.4. Only particle jets with p_{T} >15 GeV and $|\eta|$ <2.5 are considered further. Hadrons are then matched to the particle jet if they are within a distance of $\Delta R < 0.4$. In case of an ambiguity between multiple jets the closest one is matched.

When the hadrons of the truth level jets are identified, $t\bar{t}$ events are categorized based on the flavor of additional jets, excluding jets from decays of the top quarks:

- $t\bar{t}$+≥ 1b - at least one additional jet containing a B hadron
- $t\bar{t}$+≥ 1c - no additional jet containing a B hadron and at least one containing a D hadron
- $t\bar{t}$+light - remaining events

With this classification, the $t\bar{t}b\bar{b}$ process falls under the $t\bar{t}$+≥ 1b category. Events with only one jet containg a B hadron come mostly from the $t\bar{t}b\bar{b}$ with either one of the jets outside the selected phase-space or with the two B hadrons ending up in a single jet.

[1]Particle level refers to a state where all final state particles of the event after the parton shower and the haodronization are defined, but before the decays and propagation through the detector. Individual stages of the event generation were described in chapter 3, details on the particle level definition can found in reference [122].



(a) (b)

Figure 6.2: Comparison of the data to the standard model prediction in the region with at least 5 jets and 2 b-jets at the 85% working point (a) for number of jets n_{jet} and (b) for p_{T} of the jet with highest p_{T} in the event ($p_{\text{T}}^{\text{jet 1}}$). Only statistical uncertainties are included.

The actual contribution of the different $t\bar{t}$ categories depends on the selection and for the analysis regions it will be shown later in section 6.5.3. As was mentioned in chapter 3, production of c and b in a parton shower is suppressed due to their mass. Since light jets include u, d and s induced jets, the $t\bar{t}$+light component dominates unless a strict selection on the number of b-jets is introduced. This can be seen in figure 6.2, showing events with at least five jets and two b-jets selected at the 85% b-tagging working point. Since $t\bar{t}$ events contain at least two b-jets from the decays of the two top quarks, there is effectively no selection on the flavor of the additional jets.

Another thing to note in figure 6.2(a) is the increasingly worse agreement between the data and the MC with larger jet multiplicity. Even though the disagreement is quite large, it is still covered by the modeling uncertainties.

6.4.3 Modeling of the $t\bar{t}b\bar{b}$ process

The modeling of the $t\bar{t}b\bar{b}$ process is the primary source of uncertainties in the measurement. This section introduces in more detail the nominal sample used in the analysis and the different variations used as systematic uncertainties.

Nominal $t\bar{t}b\bar{b}$ Monte Carlo

There are multiple ways to approach the generation of $t\bar{t}b\bar{b}$ events. The analysis considers two main options:

1. Generate the $t\bar{t}b\bar{b}$ as a Matrix element.

2. Generate the $t\bar{t}$ process at the tree level[2] with the two b quarks coming exclusively from the parton shower. This will be referred to as $t\bar{t}$+jets.

The nominal $t\bar{t}b\bar{b}$ sample, based on the first option with the two b quarks in the matrix element, is generated with PowhegBoxRes [123] and OpenLoops [124, 125] with the PDF set `NNPDF3.0nlonf4` [126] with the mass of the two b-quarks set to $m_b = 4.95$ GeV. The factorization scale (for definition see section 3.1) is set to $\mu_f = \sum_{i=t,\bar{t},b,\bar{b},j} m_{T,i}/2$, the renormalization scale (see section 2.3) to $\mu_r = \sqrt[4]{m_{T,t} \cdot m_{T,\bar{t}} \cdot m_{T,b} \cdot m_{T,\bar{b}}}$ and the h_{damp} parameter[3] to $h_{\text{damp}} = \sum_{i=t,\bar{t},b,\bar{b},j} E_{T,i}/2$.

The nominal $t\bar{t}$+jets sample with additional b quarks coming from the parton shower only is generated as $t\bar{t}$ at the NLO using PowhegBox [46, 53]. Its h_{damp} parameter is at $h_{\text{damp}} = 1.5 m_{\text{top}}$ and its scales are set to $\mu_r = \mu_f = \sqrt{m_{top}^2 + p_T^2}$. Both generators use Pythia8 [115] as the shower generator.

One additional difference between the two samples is the implementation of the PDFs. The $t\bar{t}$+jets sample uses a 5-flavor scheme (5FS), which included all quarks with exception of the top quark in the PDF. On the other hand, the $t\bar{t}b\bar{b}$ sample includes massive b-quarks and as such they are excluded from the PDF in a 4-flavor scheme (4FS)[126].

Given that the $t\bar{t}$+jets sample does not account properly for mass of the b quarks and the fact that the parton shower is only an approximation of the additional emission. the $t\bar{t}b\bar{b}$ sample should provide more accurate predictions. However, for both samples a better agreement with the data can be achieved by tuning the parameters of the model and the performance of both samples was studied in comparison with the measured data. This is summarized in appendix B, and it was found that the $t\bar{t}b\bar{b}$ sample shows a better agreement with the data.

The $t\bar{t}b\bar{b}$ sample will be designated as Powheg +Pythia8 $t\bar{t}b\bar{b}$ (Pow+Py8 $t\bar{t}b\bar{b}$), while the $t\bar{t}$+jets sample is called PowhegBox+Pythia8 $t\bar{t}$+jets (Pow+Py8 $t\bar{t}$+jets). The next two section introduce sources of uncertainties coming from variations of model parameters and from comparison to alternative generators. The systematic uncertainties of the $t\bar{t}+\geq 1b$ component are also summarized in table 6.3.

Parameter variations

In order to determine uncertainties from tuning of the various parameters of the nominal model, their variations are produced, divided based on their physics interpretation:

Initial State Radiation (ISR): The initial state radiation systematic uncertainty comes from two sources. First, there is a variation of the factorization and renormalization scales (μ_f and μ_r, respectively) of the matrix element, parameters described previously in sections 2.3 and 3.1. They, among other things, affect properties of the additional gluon emission included in the matrix element. One variation corresponds to a half of the nominal scale and the other to a double of the nominal value. The other component of the systematic comes from a variation of the α_s^{PS} parameter of the ISR in the Pythia8 A14 Tune[118].

Final State Radiation (FSR): The final state radiation systematic uncertainty comes from a variation of the α_s^{PS} parameter of the FSR Pythia parton shower[118].

[2] Quite often the $t\bar{t}$ process is generated at NLO, so e.g. additional gluon emission can be part of the matrix element at LO, but not the subsequent splitting into a pair of (b-)quarks

[3] h_{damp} is a parameter of the Powheg NLO matching model, which controls the momentum of the first additional gluon emission.

Sample	Variation	Generator	Comments
$t\bar{t}+\geq 1b$	nominal	Pow+Py8 $t\bar{t}b\bar{b}$	
	ISR	Pow+Py8 $t\bar{t}b\bar{b}$	μ_r, μ_f and α_s^{ISR} parameter variation
	FSR	Pow+Py8 $t\bar{t}b\bar{b}$	α_s^{FSR} parameter variation
	PS&had	Pow+Her7 $t\bar{t}$+jets	with respect to Pow+Py8 $t\bar{t}$+jets
	NLO match	MG+Py8 $t\bar{t}$+jets	with respect to Pow+Py8 $t\bar{t}$+jets
$t\bar{t}H$	nominal	Pow+Py8 $t\bar{t}H$	
	ISR	Pow+Py8 $t\bar{t}H$	μ_r, μ_f and α_s^{ISR} parameter variation
	FSR	Pow+Py8 $t\bar{t}H$	α_s^{FSR} parameter variation
	PS&had	Pow+Her7 $t\bar{t}H$	
	NLO match	MG+Py8 $t\bar{t}H$	
$t\bar{t}+\geq 1c$ + $t\bar{t}$+light	nominal	Pow+Py8 $t\bar{t}$+jets	
	ISR	Pow+Py8 $t\bar{t}$+jets	μ_r, μ_f and α_s^{ISR} parameter variation
	FSR	Pow+Py8 $t\bar{t}$+jets	α_s^{FSR} parameter variation
	PS&had	Pow+Her7 $t\bar{t}$+jets	
	NLO match	MG+Py8 $t\bar{t}$+jets	

Table 6.3: Generators used for the nominal and systematic variations of the $t\bar{t}H$ signal and $t\bar{t}$ backgrounds. When two generators are listed, the first one is responsible for the Matrix element and NLO matching and the second one for the parton shower and hadronization.

The reason why variation of the matrix element is only a part of the ISR and not the FSR is because gluon emission from a top-quark is suppressed due to the large mass of the quark and in PowhegBox it is not a part of the matrix element. Due to this the impact of the ISR variation is larger than the FSR.

Parameter variations are internally provided as event weights of the nominal sample. This means that they are statistically correlated to the nominal sample, a fact which becomes important later on when studying impact of statistical fluctuations.

In addition, variations of the PDFs (see section 3.1) in the matrix element were studied but their impact was found to be negligible and is omitted.

Alternative generators

Aside from variations of parameters in the chosen nominal model, it is important to consider alternative models as well. Currently available alternative samples are only produced as the $t\bar{t}$+jets samples (with the additional b quarks produced in the parton shower).

Comparing these variations to our nominal $t\bar{t}b\bar{b}$ would lead to a double counting of the difference between the $t\bar{t}b\bar{b}$ coming from Matrix element and Parton shower. Instead, these variations are computed with respect to the $t\bar{t}+\geq 1b$ sub-component of the PowhegBox+Pythia8 $t\bar{t}$+jets sample and then applied as a systematic variation to the nominal PowhegBoxRes+Pythia8 $t\bar{t}b\bar{b}$ sample.

Systematic variations derived through a comparison of two samples are often called two-point systematics. In the case of the $t\bar{t}H(b\bar{b})$ analysis, two such uncertainties are considered:

Parton shower & hadronization (PS&had): The nominal sample uses Pythia8 for the parton shower and hadronization. The systematic on this model is assessed by comparing the nominal sample to an alternative which uses Herwig7 [40, 119] instead, designated as Powheg +Herwig7 (Pow+Her7) $t\bar{t}$+jets. The difference in the parton shower between the two samples is mainly in the choice of the ordering variable of the

shower (see section 3.3). The hadronization also relies on a different approach (string vs. cluster model), described previously in section 3.4.

NLO matching: As was mentioned before in section 3.3, there is an overlap between the NLO matrix element and the PS, which needs to be removed to avoid double counting, which is done through an NLO matching. The analysis compares two generators: PowhegBox and MadGraph5_aMC@NLO[127], which use the Powheg[46] and MC@NLO [47] models, respectively, to match the shower to the NLO matrix element. The alternative sample is called MadGraph5_aMC@NLO+Pythia8 (MG+Py8) $t\bar{t}$+jets.

Using systematic variations derived with the $t\bar{t}$+jets samples means that the two additional b quarks are generated in the parton shower and their production is thus directly affected by these systematics. This means these uncertainties are overestimated compared to the case where the two b quarks are directly part of the matrix element at the NLO.

In addition to the samples used to derive the systematic variations, another $t\bar{t}b\bar{b}$ sample is used, produced with Sherpa 2.2.1[43] interfaced with OpenLoops, with the two b quarks directly in the matrix element. It uses parton shower and hadronization models implemented directly in the Sherpa generator. This sample is used for some tests of the nominal model, but sufficient statistics was not available and fluctuations were too large to derive systematic variations in part of the analyzed phase-space.

6.4.4 Signal modeling

The nominal $t\bar{t}H$ sample is PowhegBox+Pythia8[46, 53, 128, 129] generated at NLO with `NNPDF3.0nlo` PDF[130] and with h_{damp} set to $\sqrt[3]{m_{T,t} \cdot m_{T,\bar{t}} \cdot m_{T,H}}$. It uses Pythia8 for the parton shower and hadronization.

There are four modeling uncertainties, which are similar to those of the $t\bar{t}b\bar{b}$ process, the first two being variations of internal parameters:

- ISR: The initial state radiation systematic uncertainty comes from two sources, variation of the scale of the matrix element (μ_f, μ_r) and of α_s^{PS} of the ISR in the parton shower[118].
- FSR: The final state radiation systematic comes from variation of the α_s^{PS} parameter of the QCD emission in the FSR[118].

These variations are also produced through internal weights of the nominal sample. The other two are based on a comparison to alternative generators:

- PS&had: Comparison of the nominal PowhegBox+Pythia8 sample to a sample showered with PowhegBox+Herwig7.
- NLO match: Comparison of the nominal PowhegBox+Pythia8 sample to a MadGraph5_aMC@NLO+Pythia8 alternative with a different NLO matching.

All systematic variations can be also found in table 6.3.

6.4.5 Remaining $t\bar{t}$+jets subcomponents

Though the $t\bar{t}+\geq 1b$ component of the $t\bar{t}$+jets process is the primary background, there is still a non-negligible contribution of the other $t\bar{t}$+jets components ($t\bar{t}+\geq 1c$ and $t\bar{t}$+light). They are generated as part of the $t\bar{t}$+jets PowhegBox+Pythia8 sample described in section 6.4.3 with the $t\bar{t}+\geq 1c$ and $t\bar{t}$+light heavy flavor selection.

The systematics, summarized in table 6.3, are then similar to the signal setup, where the ISR and FSR systematic uncertainties are again variations of internal weights. The PS&had systematic is defined through a comparison to Herwig7 and the NLO match to MadGraph5_aMC@NLO.

6.4.6 Small backgrounds

Remaining backgrounds are listed in table 6.4, divided into several categories. Their contribution and impact is small, so a detailed description of the samples is omitted, but their implementation did not change significantly from reference[8].

The first category is designated as $t\bar{t}$+light,$t\bar{t}t\bar{t}$,tH, which in addition to the $t\bar{t}$+light mentioned previously contains also events with 4 top quarks or tH in the final state. The three processes are joined together to avoid empty bins in the measured distributions.

The second group of background processes are $t\bar{t}V$, which refer to $t\bar{t}$ final state with one additional W or Z boson in the final state. Due to their low cross-section they have only a small yield in the analyzed regions.

The last category, called simply *other*, combines the remaining samples which individually contributed a negligible amount to the total yield. It contains final states with a single top quark in the final state (*single top*), which can be further associated with weak bosons and jets (tZq, tWZ). It also contains events with no top quarks in the final states, mainly weak bosons produced together with jets (V+jets) and production of two boson (diboson).

Category	Sample	Generator	Comments
$t\bar{t}$+light,$t\bar{t}t\bar{t}$,tH	$t\bar{t}t\bar{t}$	MadGraph5_aMC@NLO+Pythia8	
	tH	MadGraph5_aMC@NLO+Pythia8	-
$t\bar{t}V$	$t\bar{t}Z$	MadGraph5_aMC@NLO+Pythia8	-
	$t\bar{t}W$	MadGraph5_aMC@NLO+Pythia8	-
Other	Single top	PowhegBox+Pythia8	s/t-channel and Wt
	tZq	MadGraph5_aMC@NLO+Pythia8	-
	tWZ	MadGraph5_aMC@NLO+Pythia8	-
	V+jets	Sherpa	W+jets and Z+jets
	Diboson	Sherpa	-

Table 6.4: Smaller backgrounds and their nominal generator, divided in several categories used later on in the profile likelihood fit. When two generators are listed, the first one is responsible for the Matrix element and NLO matching and the second one for the parton shower and hadronization.

6.5 Event selection

An example of a Feynman diagram of the single lepton $t\bar{t}H(b\bar{b})$ final state, displayed in figure 6.3, contains four b-quarks (2 from the top decays and 2 from the Higgs boson), two additional quarks from a decay of the hadronic W and a single charged lepton. The six quarks in the final state produce jets, four of which can be tagged as b-jets. This section describes how events with these properties are selected.

6.5.1 Reconstructed object definition

Reconstructed electrons, presented in section 5.3.1, are required to have a p_{T} larger than 10 GeV and $|\eta|$ smaller than 2.47 with the barrel–endcap transition region of the calorimeter

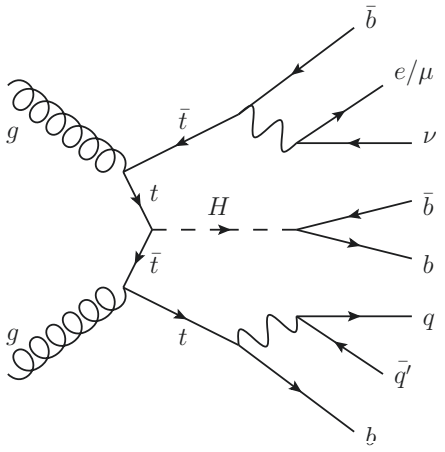


Figure 6.3: An example of a Feynman diagram of the $t\bar{t}H(b\bar{b})$ single lepton channel[28].

($1.37 < |\eta| < 1.52$) excluded. To ensure that only prompt electrons are selected, they have to satisfy the *Tight* identification criteria and the *Gradient* isolation (described in section 5.3.1). The track of the electron has to be in proximity of the primary vertex, which is ensured by a requirement on the longitudinal impact parameter $|z_0| < 0.5$ mm and on the significance of the transverse impact parameter $|d_0|/\sigma_{d_0} < 5$.

Muon candidates (see section 5.3.2) have the same p_{T} requirement of 10 GeV but a slightly larger η acceptance with $|\eta| < 2.5$. The longitudinal requirement is also the same ($|z_0| < 0.5$ mm) but the transverse is slightly tighter ($|d_0|/\sigma_{d_0} < 3$). Muons have to further pass the *Medium* identification and the *FixedCutTightTrackOnly* isolation (described previously in section 5.3.2).

The kinematic selection of the hadronically decaying tau lepton candidates requires $p_{\mathrm{T}} > 25$ GeV and $|\eta| <$2.5. Further, they have to pass the *Medium* τ-identification working point (see section 5.6). Tau leptons decaying leptonically produce electrons and muons which are not distinguished from other prompt leptons.

The analysis uses jets reconstructed using the anti-k_t algorithm with a radius 0.4 (as described previously in 5.4) and with kinematic requirements $p_{\mathrm{T}} > 25$ GeV and $|\eta| < 2.5$. Additional quality criteria described in reference [131] are used to remove jets from non-collision sources and calorimeter noise. A selection on the jet vertex tagger (described in section 5.4.4) is applied for jets satisfying $p_{\mathrm{T}} < 60$ GeV and $|\eta| < 2.4$ to reduce the contribution from pile-up jets.

Since a final state with a large number of b quarks is studied, b-tagging and b-jets identification (see section 5.5) are an essential part of the analysis. The analysis mainly considers the 60% and 70% b-tagging efficiency working points. This allows definition of more varied analysis regions, either with a higher contribution of signal or control regions which help to constrain the backgrounds.

Finally, an overlap removal, as described in section 5.7, is applied to each event.

6.5.2 Single lepton region definition

All events in the $t\bar{t}H(b\bar{b})$ analysis are required to pass the triggers shown in table 6.2 and have to have all relevant detector systems operational. Furthermore, they have to contain a reconstructed primary vertex as described in section 5.2.2. In the single lepton channel, the events are further required to have exactly one electron or muon with p_{T}

>27 GeV[4] and no additional lepton with $p_{\mathrm{T}} > 10$ GeV. Additionally, to remove overlap with other $t\bar{t}H$ analyses, a veto on two or more tau leptons with hadronic decay is used. This mainly concers overlap with the multi-lepton final state[31, 32] which considers region with hadronic tau leptons in the final state.

Based on the final state of the $t\bar{t}H(b\bar{b})$, the signal and the $t\bar{t}b\bar{b}$ background is expected to contain at least six jets. However, one can also get a relatively clean $t\bar{t}H(b\bar{b})$ $+t\bar{t}b\bar{b}$ sample with only five jets in the final state with a sufficient selection on the number of b-jets, which slightly increases the available statistics (as will be shown in section 6.5.3). This creates two jet-multiplicity classes, one with exactly 5 jets and one with at least 6 jets. They form separate regions and help to mitigate the mis-modeling of the variable number of jets (n_{jet}, shown previously in figure 6.2(a)). The 5-jets option contributes only slightly to the sensitivity due to lower number of events and is considered for control regions to better constrain the background model.

Furthermore, there are several working points of the b-tagging algorithm (described in section 5.5), which can be used to separate the signal and the different backgrounds. Starting with the tightest working point, one can derive relatively pure $t\bar{t}H/t\bar{t}b\bar{b}$ regions by selecting at least 4 b-jets at the 60% working point. Such regions are labeled $\geq 4b$ *hi*. However, they still contain a small amount of $t\bar{t}+\geq 1c$ and $t\bar{t}$+light. To get a better handle on these additional backgrounds and to get better statistics for the $t\bar{t}+\geq 1b$ final state, looser regions in the b-tagging are defined. They require 4 b-jets at the 70% working point (but veto the previously defined region with the 60% working point), resulting in events with a label $\geq 4b$ *lo*. By combining the two jet multiplicity categories and the two b-tagging categories one arrives to four analysis regions, summarized in table 6.5.

Region	n_{lepton}	n_{jet}	$n_{\text{b-jet}}$ @60%	$n_{\text{b-jet}}$ @70%
$\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$	$= 1$	≥ 6	≥ 4	≥ 4
$\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$			< 4	
$\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$		$= 5$	≥ 4	
$\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$			< 4	

Table 6.5: The definitions of the single lepton analysis regions, based on the number of jets n_{jet}, and the number of b-tagged jets $n_{\text{b-jet}}$ using the 60% and 70% working points. SR refers to signal regions and CR to control regions.

6.5.3 Region composition

The background composition of the four analysis regions can be found in figure 6.4(a), where different fractions of the $t\bar{t}$+jets components between the regions with different b-tagging requirements can be observed. The difference due to the number of jets is minimal, since the 5 jet regions are dominated by the $t\bar{t}$+jets process where one of the jets from the W is not reconstructed in the acceptance of the detector. Since all $t\bar{t}$+jets components contain a W from the top, the missing jet does not affect the relative contribution of the different $t\bar{t}$+jets components.

The signal to background ratio (S/B), where S and B are number of signal and background events, respectively, and the approximation of a statistical significance of the

[4]The 27 GeV threshold is chosen to reflect the trigger which has a 26 GeV selection. The 1 GeV difference is there to remove a region where the efficiency of the trigger is not well measured, resulting in large uncertainties.

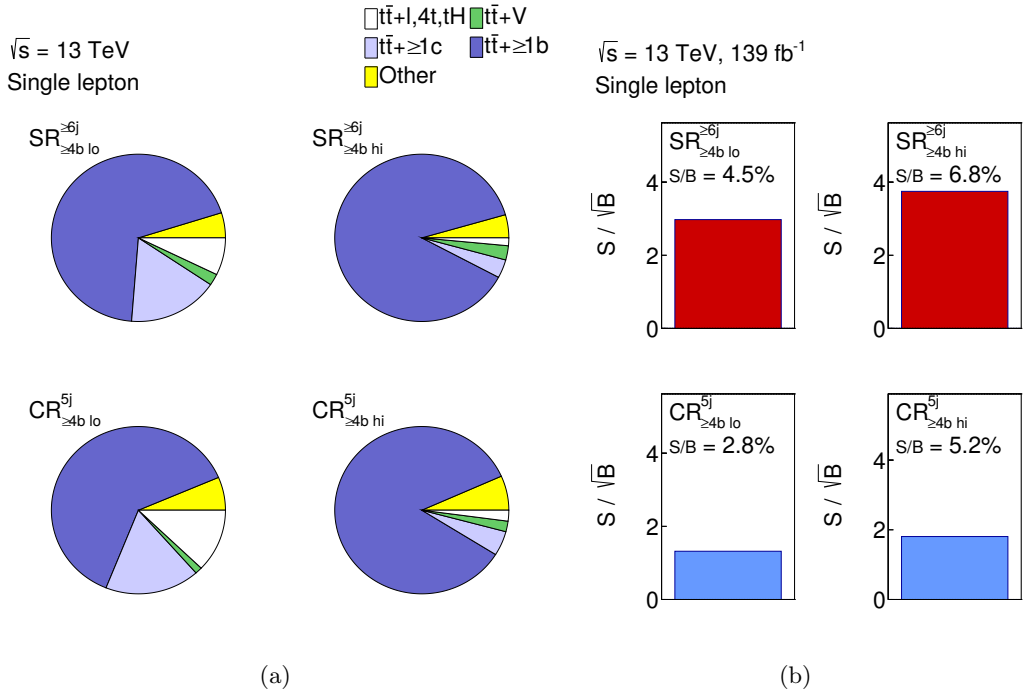


Figure 6.4: The background composition (a) and signal over background ratio and statistical significance (b) of the four analysis regions.

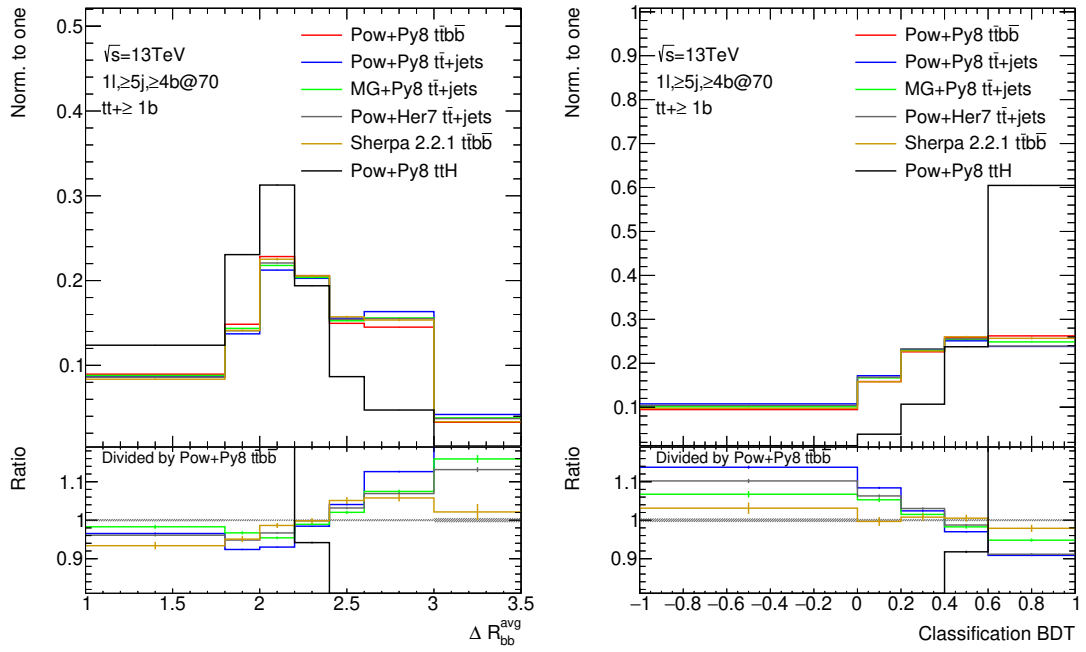
signal ($S/\sqrt{B}$) is displayed in figure 6.4(b). There, one can see a larger significance in the six jet regions compared to their five jet counterparts. In the region with the tighter b-tagging requirement ($\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$) the statistical significance is almost 4σ.

6.6 Multi variate algorithms and variables used in the fit

The analysis takes advantage of several multi-variate algorithms (MVA) to connect the reconstructed objects to the underlying particles and to better distinguish between the signal and background events. A short description of the main multivariate technique used in the analysis, the Boosted Decission Tree (BDT), can be found in appendix A. To separate the signal from the background, the analysis uses an MVA designated as Classification BDT, which is described in more detail in the appendix A.3. The output values of the BDT lie between -1 and 1, larger values being more signal like.

The analysis uses a profile likelihood template fit to extract the signal (see chapter 7). The Classification BDT is used in the fit of the 6-jet regions. One of the inputs of the Classification BDT is the average ΔR for all pairs of b-tagged jets ($\Delta R^{\mathrm{avg}}_{bb}$). This variable is used in the 5-jet regions and helps to constrain the background modeling. Both variables are displayed for different Monte Carlo generators in figure 6.5. Several properties can be observed:

- Both variables provide a signal to background discrimination, which is larger for the Classification BDT as per design
- The nominal Pow+Py8 $t\bar{t}b\bar{b}$ sample is the most signal-like among the $t\bar{t}+\geq 1b$ Monte Carlo samples



(a) (b)

Figure 6.5: Distributions of the nominal $t\bar{t}H$ and $t\bar{t}b\bar{b}$ sample, and alternative generators falling under the $t\bar{t}+\geq 1b$ classification as function of (a) $\Delta R_{bb}^{\text{avg}}$ and (b) the Classification BDT. The generators follow designation introduced in section 6.4.5. All distributions are normalized to unity.

- $\Delta R_{bb}^{\text{avg}}$ provides slightly larger discrimination between the $t\bar{t}+\geq 1b$ samples than the BDT, especially for the nominal and the Sherpa 2.2.1 $t\bar{t}b\bar{b}$ sample

6.7 Techniques for input preparation

The statistical model uses Monte Carlo distribution to build templates, which are further used in the profile likelihood fit. Important step in the construction of the fit model is the choice of the binning of these templates. Its definition is crucial since it affects the sensitivity and impact of the statistical fluctuations. Another important point is the practical implementation of the systematic uncertainties, where various smoothing techniques are used to minimize impact of statistical fluctuations of the variations.

6.7.1 Binning

The binning is an important part of the construction of the analysis regions for the likelihood fits. Fewer bins means larger statistics in each, but the trade-off is usually a lower control over the backgrounds when the systematic uncertainties are large, leading to a lower sensitivity. The binning in the analysis was optimized and chosen based on a method described in detail in reference [132]. In the $t\bar{t}H(b\bar{b})$ analysis, the number of bins for the 5-jet regions is $N_{\text{bins}} = 6$, while for the 6-jet regions it is set to $N_{\text{bins}} = 8$.

6.7.2 Smoothing

In the previous $t\bar{t}H(b\bar{b})$ measurement, statistical uncertainties of the Monte Carlo samples were one of the dominant factors impacting the significance. Their effect will be studied later in section 7.8. Here, the focus is on techniques used to minimize the impact of statistical fluctuations on the systematic uncertainties.

In principle, an unmodified distribution of the uncertainty can be used in the fit. However, statistical fluctuations of systematic variations can match (simply by chance) a fluctuation in the data, which can lead to a large impact on the fit result. One way to minimize this issue is to introduce statistical uncertainties on the systematic. However, the model then becomes significantly more complicated. Furthermore, derivation of these statistical uncertainties is not straightforward for systematic uncertainties which are correlated to the nominal sample[5].

Instead, a smoothing is applied on the templates used to derive systematics to average out the fluctuations. This can introduce a bias into the analysis which needs to be investigated to make sure its impact on the result of the analysis is not large.

Smoothing algorithms

Several smoothing methods are used in the analysis, all implemented in an ATLAS internal smoothing software package. These used in the analysis preserve the effect on normalization due to the systematic uncertainty. They are based on merging of bins and subsequent smoothing of the shape. The binning of the nominal distribution remains unchanged.

Though a larger number of methods was tested, three methods were studied in more detail as they better captured the shape of the systematic. Here, the algorithms are introduced, their performance is then described in the next section.

The first method is called Parabolic, which is implemented as described in figure 6.6. It first merges pairs of bins based on their χ^2 compatibility until the number of slope changes

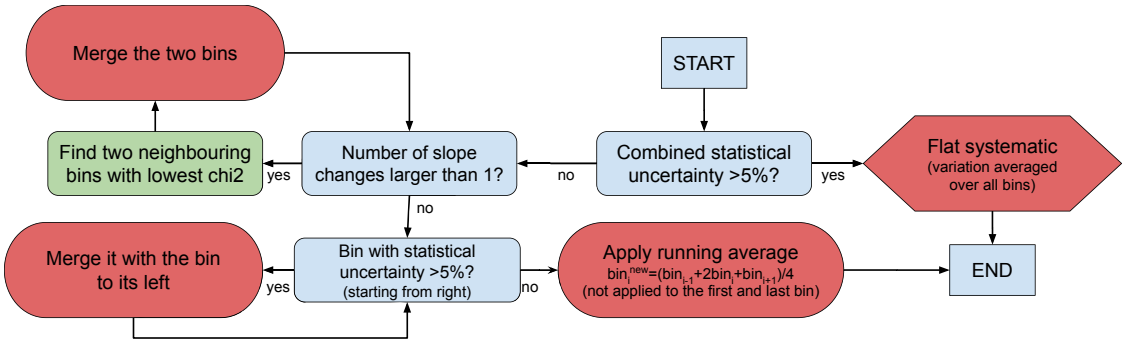


Figure 6.6: Flowchart describing the process of the Parabolic smoothing on a binned systematic.

in the distribution is smaller than two. Then bins with a large statistical uncertainty are merged to the left and the whole distribution is smoothed by a running average[6].

Two other smoothing algorithms are considered. They both use the TH1::Smooth algorithm at some stage, an internal function of the data analysis software ROOT[133]

[5]More information on how to derive statistical uncertainty of a correlated systematic is given in section 7.8.

[6]Running average means that each bin i is updated as $bin_i^{new} = (bin_{i-1} + 2bin_i + bin_{i+1})/4$ with the exception of the first and the last bin.

based on the 353QH TWICE smoothing algorithm[134]. First, there is the MAXVAR algorithm, described in figure 6.7. This method merges bins with a large statistical

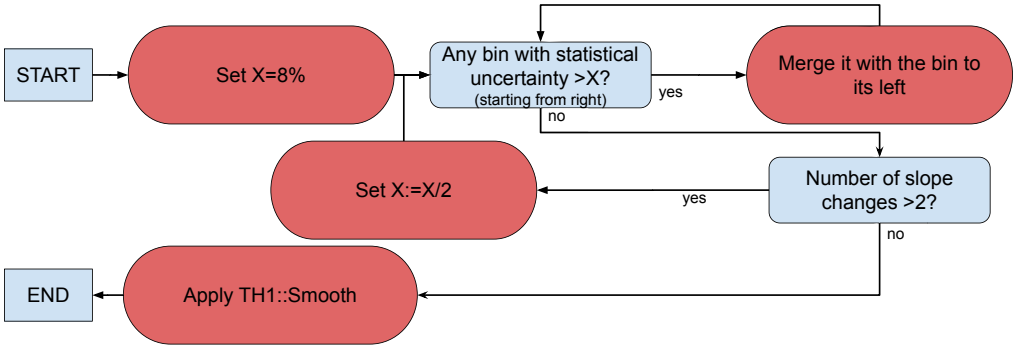


Figure 6.7: Flowchart describing the process of the MAXVAR smoothing on a binned systematic.

uncertainty until there are no more than 2 changes of slope. The TH1::Smooth algorithm is then used to smooth out the shape.

The final algorithm is called TTRES, shown in figure 6.8. It has the simplest implementation, merging neighboring bins until all pairs have χ^2 larger than one. Then, the TH1:Smooth algorithm is applied.

Performance and application of the smoothing algorithms

The differences in performance between the three algorithms can be seen in figure 6.9, in which their impact on two different systematics are shown: a small systematic with a large statistical uncertainty (one of the b-tagging uncertainties) and a larger uncertainty from modeling (the $t\bar{t}b\bar{b}$ NLO matching systematic). Two important features can be observed:

1. The PARABOLIC smoothing introduces a shape for the small systematic due to a fluctuation in one bin. Given the large statistical uncertainties of the underlying distribution, a flat shape, which is produced by MAXVAR and partially also TTRES, seems more reasonable as an approximation.

2. Both MAXVAR and TTRES fail to capture the last bin of the large systematic.

Because of the first feature it was decided to use the MAXVAR smoothing for systematic uncertainties, where the difference between bins is comparable to statistical uncertainties. This choice avoids artificial shapes and mainly concerns experimental uncertainties and modeling of small backgrounds.

The second feature was found to be related to the 353QH TWICE smoothing implemented in TH1::SMOOTH as part of the MAXVAR and TTRES, which leads to smoothed shapes which often fail to capture the first and the last bins of the distributions. This is due to a floating median (described in reference [134]) applied in the first step, which ignores content of the border bins if they differ significantly from the two bins closest to them. Since the analysis uses a classification discriminant for the fitting, the last bin is the one most sensitive to the signal and its bad modeling would have a significant impact on the result of the analysis. This is the main reason why algorithms using the TH1::SMOOTH are not applied on systematics with a large variation in the first and the last bins.

To summarize, the MAXVAR algorithm is used for all experimental uncertainties and modeling systematics of the small backgrounds (all backgrounds except for the three $t\bar{t}$+jets

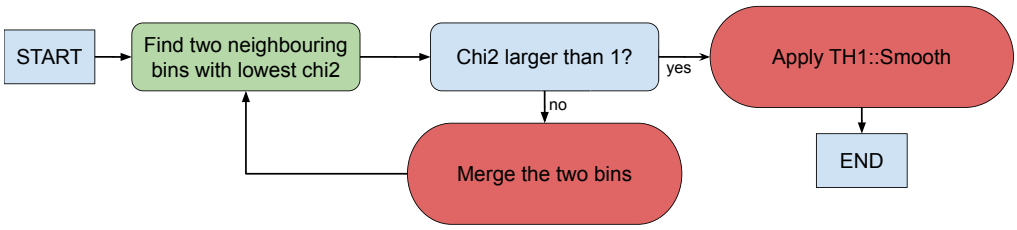


Figure 6.8: Flowchart describing the process of the TTRes smoothing on a binned systematic.

components). The smoothed distributions were checked to see that they capture the shape of the original distribution within the uncertainties. The remaining systematics (modeling of the $t\bar{t}$+jets components and of the signal) have much larger shape and because of that a larger impact on the result and the Parabolic algorithm was found to perform the best in capturing of their shape and is used as a default smoothing. For three of the modeling systematics, $t\bar{t}+\geq 1b$ NLO matching (NLO match), $t\bar{t}H$ NLO match and $t\bar{t}H$ PS, only the Parabolic is able to capture the shape properly and is therefore used exclusively (one can see it back in the figure 6.9 for $t\bar{t}+\geq 1b$ NLO matching). Using the other smoothing methods would mean underestimating this systematic variation significantly. For the remaining modeling systematics, the TTRes is used as an alternative to study the impact of the smoothing on the fit. This is shown later in section 8.3.3.

6.7.3 Symmetrization of two-sided systematic uncertanties

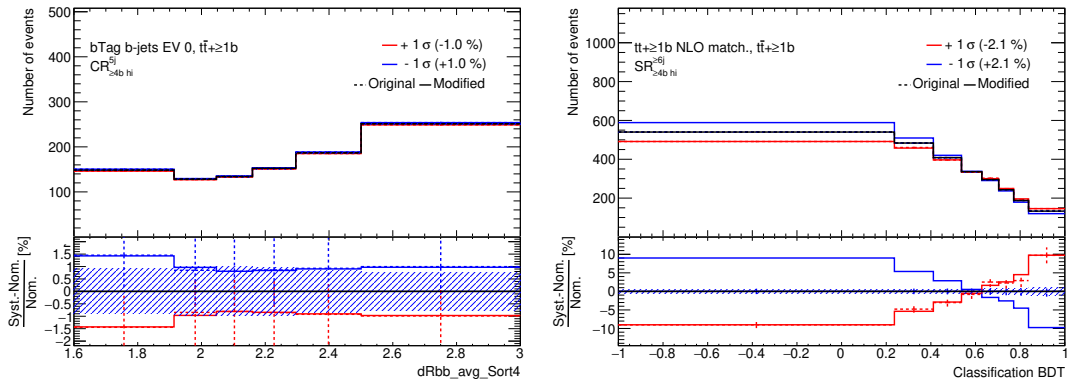
A systematic uncertainty can either increase or decrease the yield, designated with an *up* and a *down* label. However, in many cases the direction of the systematic changes for different values of an observable[7]. While the up and down label is arbitrary it is important to keep the proper correlation between the bins.

Even if the systematic uncertainty is symmetric, the input distribution does not have to necessarily reflect that because of statistical fluctuations. To minimize the statistical impact, one can calculate new variation as an average of absolute variations, creating symmetric uncertainties. This reduces the impact of statistical fluctuations without changing the underlying systematic. In case of e.g. two-point systematics, only one side of the variation is available. Such variations are symmetrized.

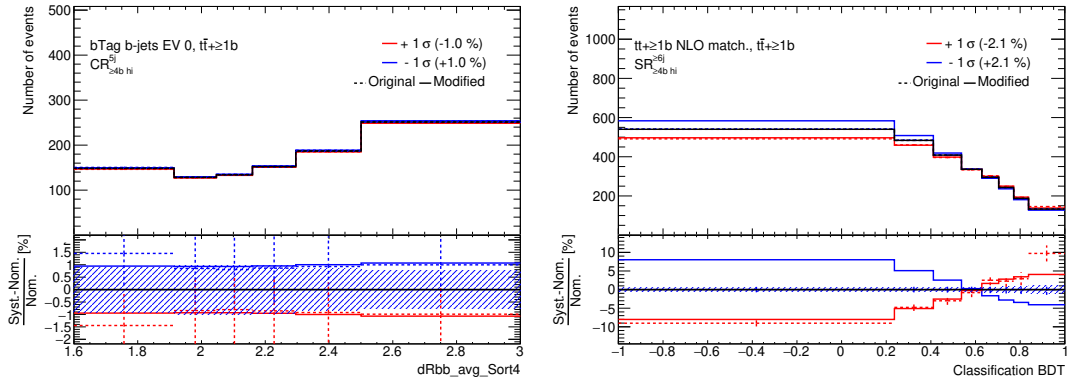
6.7.4 Factorization of the normalization

The fraction of $t\bar{t}+\geq 1b$ events in the $t\bar{t}$+jets process is a badly modeled property. For this reason, the normalization of $t\bar{t}b\bar{b}$ is left free-floating in the fit and is designated as $k(t\bar{t}+\geq 1b)$. Because of that the normalization effect of $t\bar{t}b\bar{b}$ systematic variations does not impact the results of the fit. However, if their normalization effect is large, these systematics will be highly correlated to the $k(t\bar{t}+\geq 1b)$, which makes it difficult to disentangle their effects in the fit. For this reason, the $t\bar{t}+\geq 1b$ systematic variations are renormalized to not change the normalization inclusively across all analysis regions (including both the single lepton regions and the dilepton regions explored in chapter 8). This means that they can have a normalization effect when used in only one of the channels and in individual regions. The procedure is used on all $t\bar{t}+\geq 1b$ systematics mentioned hereafter.

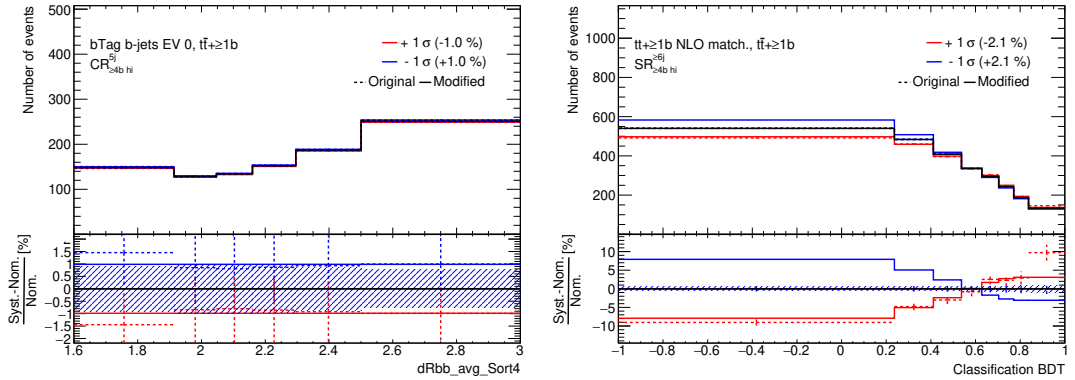
[7]Several examples of this will be presented in the rest of the chapter.



(a) Parabolic

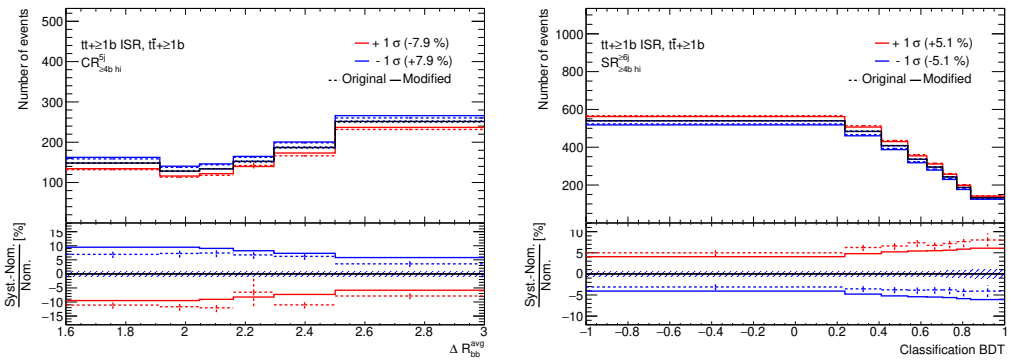


(b) MaxVar



(c) TTRes

Figure 6.9: The effect of various smoothing methods on a $t\bar{t}+\geq 1b$ b-tagging systematic in the $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ region (left) and on the $t\bar{t}+\geq 1b$ NLO match in the $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ region (right). Each figure shows the original distribution used as an input and the modified distribution with the smoothing and symmetrization applied. The two systematics are described in more detail in section 6.8.



(a) (b)

Figure 6.10: Distributions of the $t\bar{t}+\geq 1b$ ISR systematic in two analysis regions: $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (a) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (b). *Original* refers to the raw input distribution, *modified* is the distribution after symmetrization and smoothing.

6.8 Shapes of major systematic uncertainties

The various sources of systematic uncertainties used in the analysis were introduced in previous chapters, coming from the modeling of different processes or from the various objects reconstructed and identified by the detector. This section presents a selection of systematic uncertainties, which play a large role in the analysis.

6.8.1 $t\bar{t}b\bar{b}$ modeling systematics

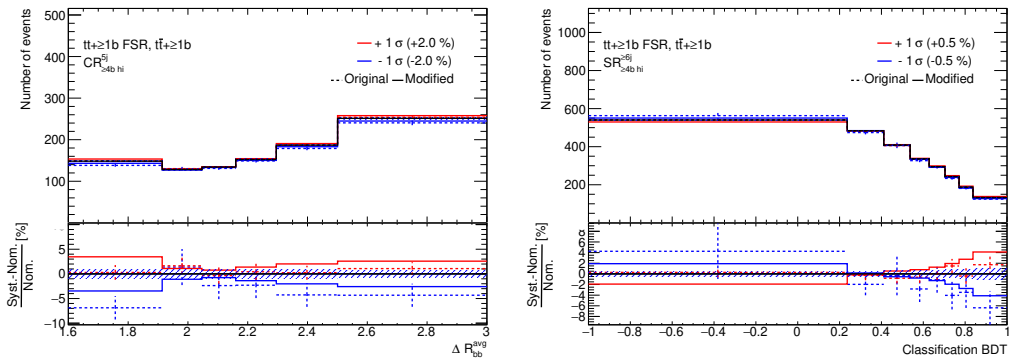
The modeling of the $t\bar{t}b\bar{b}$ process is the main limiting factor of the $t\bar{t}H(b\bar{b})$ analysis. All sources of the $t\bar{t}b\bar{b}$ modeling systematic uncertainties were described in section 6.4.3. They all use the symmetrization and the Parabolic smoothing described previously. A few examples of their distributions will be shown in this section, all of them can be found in appendix C.

The **Initial State Radiation (ISR)** systematic is displayed in figure 6.10. There is a slight asymmetry between the up and down component in the original distributions before the application of the smoothing and symmetrization. This raises a question whether the symmetrization should be used in such case and its effect was studied in the single lepton channel, where the statistics is high enough not to deem symmetrization necessary. The effect on the fit result was found to be negligible and therefore the smoothing is applied anyway to improve precision of the systematic in the dilepton channel, where the statistical fluctuations are higher.

As was described in section 6.4.3, the ISR systematic represents variation in the production of additional jets in both the matrix element and the parton showers. This means it significantly affects the jet multiplicity, which can be seen by an opposite direction of the systematic in the 5-jet and the 6-jet regions. Otherwise the shape of the systematic is relatively flat.

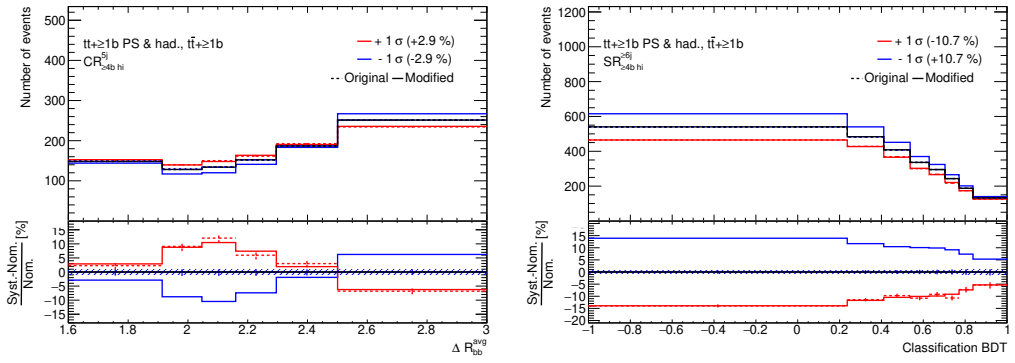
The **Final State Radiation (FSR)** variations, displayed in figure 6.11, have the largest statistical fluctuations. This is due to large variations of the weights used to derive this uncertainty. The impact of these fluctuations on the result of the fit was tested using Monte Carlo toys and was found to be negligible, as will be described later in section 7.8.

The first two-point systematic $t\bar{t}+\geq 1b$ **Parton shower & hadronization (PS&had)**



(a) (b)

Figure 6.11: Distributions of the $t\bar{t}+\geq 1b$ FSR systematic in two analysis regions: $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (a) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (b). *Original* refers to the raw input distribution, *modified* is the distribution after symmetrization and smoothing.



(a) (b)

Figure 6.12: Distributions of the $t\bar{t}+\geq 1b$ parton shower systematic in two analysis regions: $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (a) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (b). *Original* refers to the raw input distribution, *modified* is the distribution after symmetrization and smoothing.

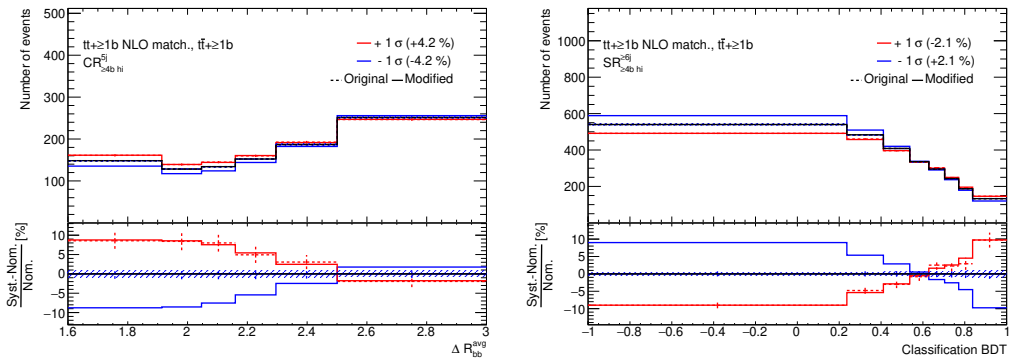
is displayed in figure 6.12. It shows a large dependence on the $\Delta R^{\mathrm{avg}}_{bb}$ variable, while it is relatively flat in the BDT.

Finally, the **NLO matching (NLO match)** systematic is shown in figure 6.13. It has a large shape with difference up to 20% in the $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ between the first and the last bin. It will be shown later that this systematic is strongly correlated to the signal normalization, making it a limiting factor of the analysis.

6.8.2 $t\bar{t}H$ modeling uncertainties

Distributions of the four $t\bar{t}H$ modeling systematics, described previously in section 6.4.4, are displayed in figure 6.14 for the tightest region ($\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$) where they will have the largest effect. All variations in all regions can be found in appendix C. The fluctuations of the FSR systematic are not as high as in the case of $t\bar{t}b\bar{b}$, though smoothing is still necessary. The PS&had systematic has the largest normalization effect.

In addition to the modeling systematic variations, there are uncertainties on the predicted cross-section, one coming from a variation of the QCD scale with $^{+5.8\%}_{-9.2\%}$ and one



(a) (b)

Figure 6.13: Distributions of the $t\bar{t}+\geq 1b$ NLO matching systematic in two analysis regions: $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (a) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (b). *Original* refers to the raw input distribution, *modified* is the distribution after symmetrization and smoothing.

from a variation of the PDF and α_S, which gives 3.6% uncertainty[3, 135]. In addition, there is a 2.2% uncertainty on the branching fraction of the decay of a Higgs to a pair of b quarks[3].

6.8.3 Other $t\bar{t}$+jets uncertainties

The modeling uncertainties of the other $t\bar{t}$+jets sub-components were described in 6.4.5. On top of these, there are two normalization uncertainties. Similarly to the $t\bar{t}+\geq 1b$, the fraction of $t\bar{t}+\geq 1c$ events is a badly modeled property. The previous iteration of the analysis reported a normalization factor of around 1.6[8]. To cover this value, a 100% uncertainty is put on the $t\bar{t}+\geq 1c$ normalization. For the $t\bar{t}$+light, which dominates the $t\bar{t}$+jets production, the cross-section is better understood and its uncertainty is only 6%[136].

6.8.4 Experimental systematic uncertainties

Unlike the modeling uncertainties, which only affect a single sample, the experimental uncertainties, which describe the performance of the detector and of the reconstruction and calibration procedures (see chapter 5), are correlated across all samples.

Since the properties of the colliding bunches are not understood perfectly, an 1.7% uncertainty on the luminosity[61] and additional uncertainty on the pile-up modeling[137] are included in the analysis. There are also several uncertainties related to the muon and electron reconstruction and identification but their impact was found negligible.

For jets, there are several components, related to the jet energy scale and jet energy resolution, correcting the jet four-momentum. Furthermore, there are uncertainties on the jet vertex tagger and the E_{T}^{miss} reconstruction.

B-tagging uncertainties are divided into three categories based on the type of the jet:

- **bTag b-jets:** uncertainty on the efficiency of jets coming from the decay of a B hadron.
- **bTag c-jets:** uncertainty on the mis-tag rate of jets coming from the decay of a D hadron.
- **bTag light-jets:** uncertainty on the mis-tag rate of jets coming from other sources.

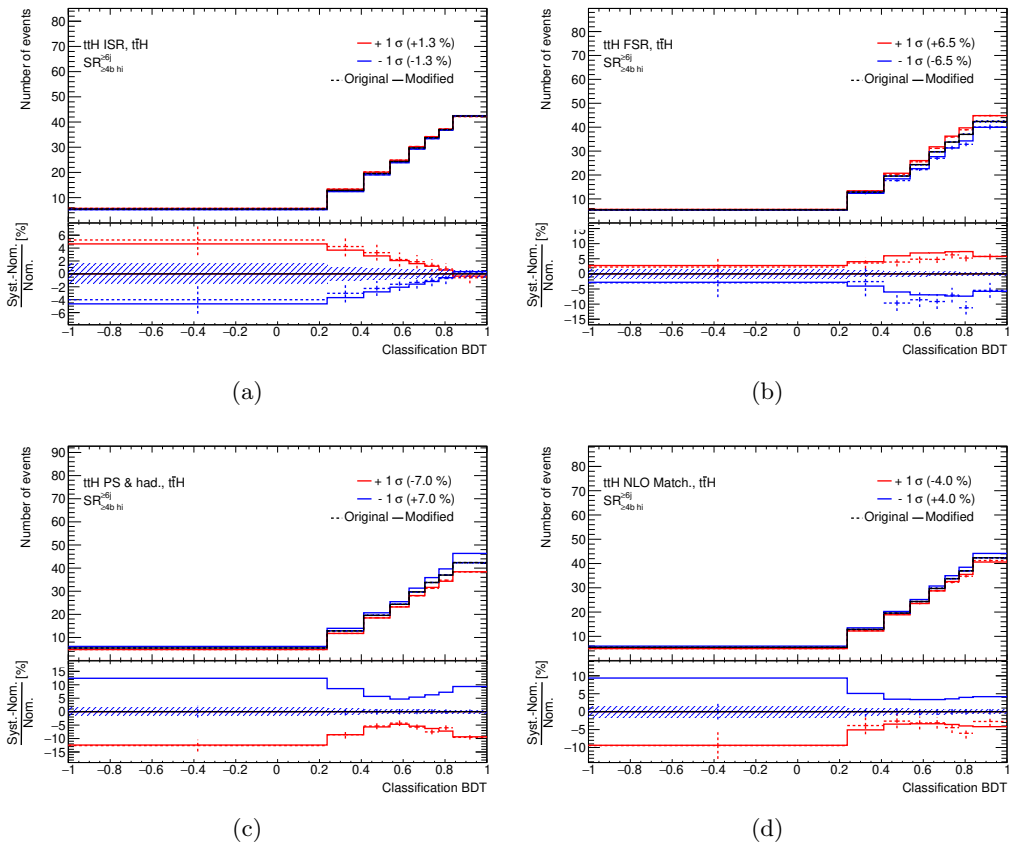


Figure 6.14: Distributions of the $t\bar{t}H$ modeling systematics - ISR (a), FSR (b), Parton Shower (c) and NLO matching (d) - in the $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ region. *Original* refers to the raw input distribution, *modified* is thedistribution after symmetrization and smoothing.

These categories do not correspond to the $t\bar{t}$+jets categories introduced previously: the $t\bar{t}+\geq 1b$ sub-component can still contain jets from decays of D hadrons (e.g. from $W \to c\bar{s}$ decays), while both $t\bar{t}+\geq 1c$ and $t\bar{t}$+light still include b-jets coming from decays of the top quarks. Thus, all three systematic categories contribute to all three components.

The b-tagging systematics are further divided into a large number of components, designated by EV X, with X starting at 0. The components are ordered based on the size of their impact in the phase-space they were derived in, so the order to a certain degree depends on the analyzed phase-space.

6.9 Comparison to the data

With a complete set of systematic variations, all components of the nominal model are available for the statistical analysis. For the regions later used in the fit, a comparison of the data and the prediction can be found in figure 6.15. More distributions can be found in appendix D.

The prediction underestimates the data, which is mainly connected to the bad modeling of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ fractions. In the previous iteration of the analysis the normalizations were found to be underestimated by 24% and 63% , respectively[8]. Furthermore, there is a normalization difference between the 5 jet and the 6 jet regions, which suggests that the ISR systematic, discussed previously in section 6.8.1, will play an important role

to compensate for this discrepancy. Beside the normalization effects there does not seem to be a significant difference in shape between the data and the Monte Carlo.

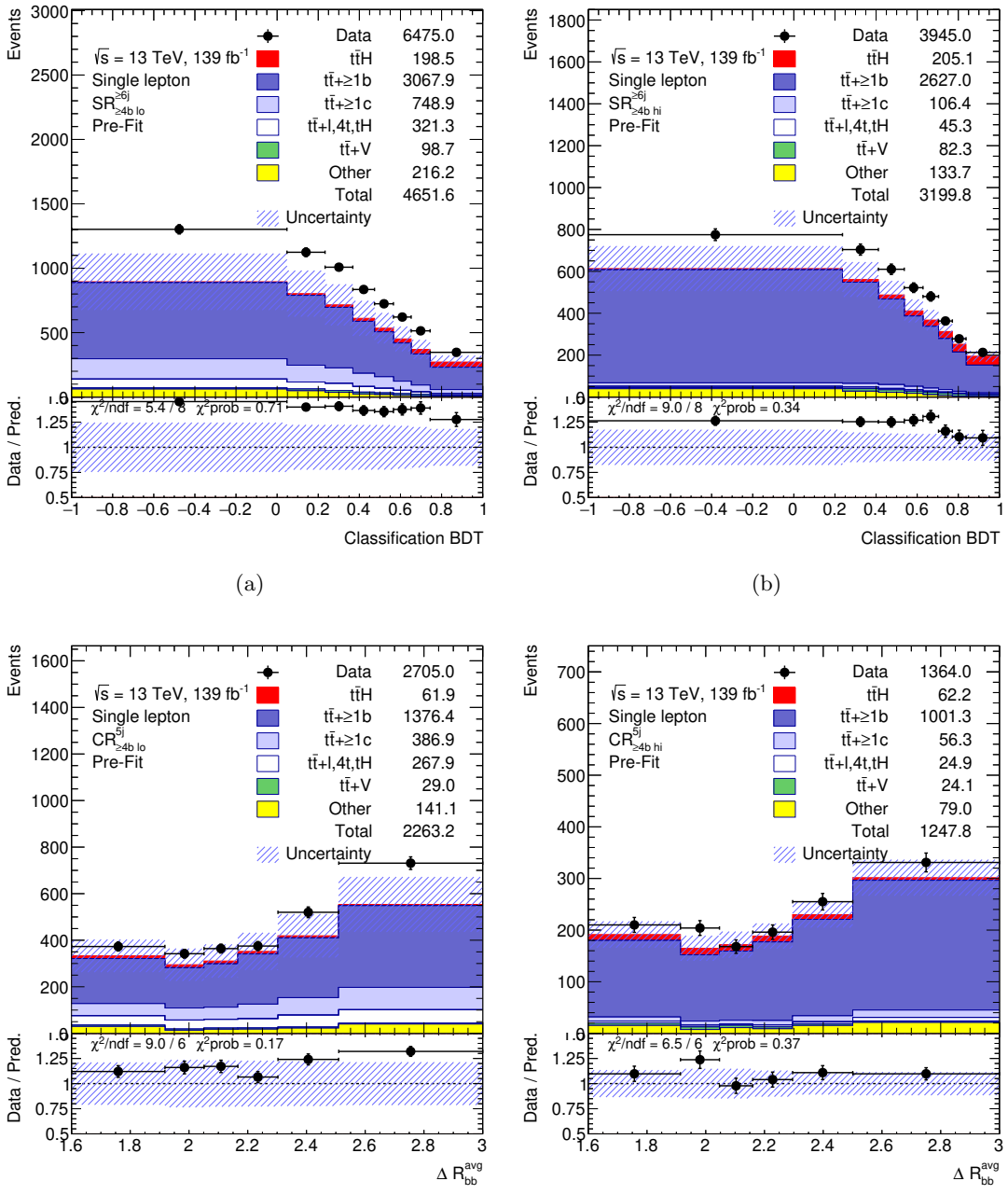


Figure 6.15: Pre-fit modeling of the four analysis regions, (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$, (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$, (c) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{lo}}$ and (d) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{hi}}$, displayed as a function of the discriminating variable used later on in the fit, $\Delta R^{\mathrm{avg}}_{bb}$ for the 5-jet regions and the classification BDT for 6-jet regions.

CHAPTER 7

Statistical analysis of the single lepton channel

The $t\bar{t}H(b\bar{b})$ analysis uses a profile likelihood fit to extract the signal strength from the data. Various studies are performed to validate the input distributions and the performance of the fit. It is important to note that most of the studies were also done in combination with a second selection channel using dilepton events, which is shown in the next chapter. It will be shown that the combination of the two channels provides additional information and can be used to better control the background .

Still, it is useful to study the single lepton channel in detail to understand the important ingredients of the analysis in a simpler environment. Furthermore, some challenges encountered in the previous analysis, like the impact of limited MC sample statistics, are studied only for the single lepton channel.

This chapter opens with a theoretical description of the profile likelihood fit method in section 7.1. Then, some of the studies performed to test the nominal model of the signal and the background, which was introduced in the previous chapter, are presented in sections 7.3 to 7.7.

Section 7.8 is dedicated to studies of the limited statistics of the Monte Carlo samples and their impact on the results. This is of special importance, since these effects were the second largest source of systematic uncertainties in the previous iteration of the analysis. Finally, results of a fit to the data are shown and compared to expectations from various studies in section 7.9.

7.1 Profile likelihood fit

The goal of the analysis is to measure the $t\bar{t}H(b\bar{b})$ process. This is done by extracting from the data the signal strength, defined as a ratio between the measured cross-section and the SM prediction: $\mu_{t\bar{t}H} = \frac{\sigma^{\text{measured}}}{\sigma^{\text{SM}}}$. To achieve this goal, a profile likelihood fit is used. The main principles behind this parameter-estimation method are introduced in the following sections.

7.1.1 Likelihood function

Any kind of data analysis works with a dataset or a sample x of measured properties, which one tries to connect to some theoretical quantities $\vec{\theta}$ (usually parameters of a model). Scientific models have to give a probability to observe some data for given values of its

parameters $P(x|\vec{\theta})$. A likelihood function is then this probability defined for given observed data x^{obs}:

$$L(\vec{\theta}) = P(x^{\text{obs}}|\vec{\theta}). \tag{7.1}$$

The likelihood can be used in a parameter estimation. It can be shown, that by finding its global maximum over all parameters the resulting values of the parameters $\hat{\vec{\theta}}$ are good estimators of their true value $\vec{\theta}^{\text{true}}$[138, 139]. The estimated parameters may be biased, but the bias disappears in the large sample limit. This method of parameter estimation is called the Maximum Likelihood method. One of its features is that it is invariant under a variable transformation. Since it looks for an extreme of the function, the derivation at that point over all parameters yields zero:

$$\frac{\partial L(\vec{\theta})}{\partial \vec{\theta}}|_{\hat{\vec{\theta}}} = 0. \tag{7.2}$$

In general, it is not possible to express the estimator $\hat{\vec{\theta}}$ from the equation 7.2 directly and computational methods are used to scan the parameter phase-space to find the maximum value of the likelihood.

In practical implementations, a logarithm of the likelihood (or log-likelihood, LL) is used instead. Since the logarithm is a monotonic function, it does not affect the parameter estimation and equation 7.2 holds even if one replaces the likelihood with the log-likelihood.

7.1.2 Variance of the parameter estimator

Knowing the value of an estimator is useless without knowledge of its uncertainty. There are several approaches how it can be determined. One approach is based on the second derivative of the likelihood, which gives an estimator of variance V_{ij} of two parameters θ_i, θ_j in the following form[138]:

$$\hat{V_{ij}} = 1/\left(-\frac{\partial^2 \log L}{\partial\theta_i\theta_j}\right)|_{\hat{\theta}_i\hat{\theta}_j}, \tag{7.3}$$

from which one can determine the uncertainty of a parameter θ_i:

$$\Delta\hat{\theta}_i = \sqrt{\hat{V_{ii}}} = 1/\sqrt{-\frac{\partial^2 \log L}{\partial\theta_i^2}}|_{\hat{\theta}_i}. \tag{7.4}$$

This method requires only computation of the second derivative at a single point, but gives a strictly symmetrical uncertainty and uses a quadratic approximation in the maximum of the log-likelihood. Such approximation only holds exactly in the large sample limit, but for a small number of events neither usually holds.

A more precise way to estimate the error is through a graphical method[138], which can give different values for the upper $\Delta\hat{\theta}_i^{\text{up}}$ and lower $\Delta\hat{\theta}_i^{\text{down}}$ uncertainty through the following relation:

$$\log L(\hat{\theta}_i + \Delta\hat{\theta}_i^{\text{up}}) = \log L(\hat{\theta}_i - \Delta\hat{\theta}_i^{\text{down}}) = \log L(\hat{\theta}) - 1/2 \tag{7.5}$$

This is illustrated with an example in figure 7.1. The analytical solution of this equation is usually not possible and one needs to scan the likelihood to find the values of the uncertainties. This makes this approach much slower compared to the previous method, where one only has to evaluate the second derivative in a single point, but more precise. In the context of the $t\bar{t}H(b\bar{b})$ measurement, the graphical method is only used to determine uncertainties of important parameters and the former method is used otherwise.

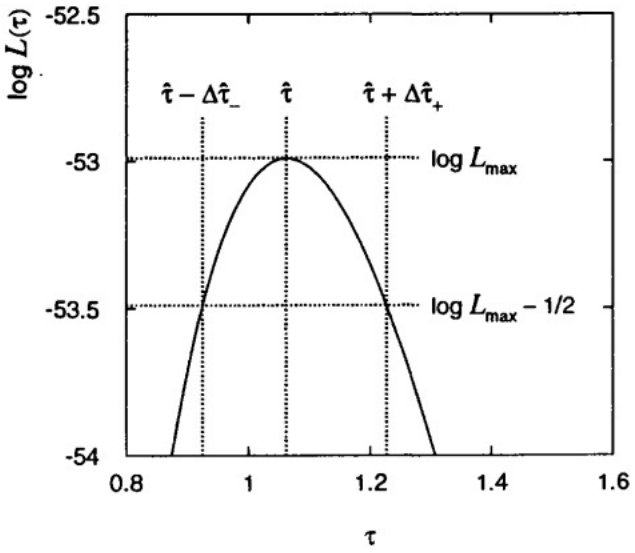


Figure 7.1: Visual depiction of the graphical method for an estimation of a parameter uncertainty, taken from [138].

7.1.3 Binned maximum likelihood

Quite often the measured data are analyzed using a binned distributions: *histograms*. This means that each bin i has a certain number of data events n_i and the theoretical prediction gives a mean number of events in each of the bins $\nu_i(\vec{\theta})$, which again depends on some underlying parameters of the model $\vec{\theta}$.

The data in each bin then follow a Poisson distribution $P_{\text{Poisson}}(n, \nu_i)$ and the total probability of observing given data is just a multiplication of the probability in each bin. This gives a log-likelihood in the following form:

$$\log L(\vec{\theta}) \propto \sum_{i}^{N_{\text{bins}}} \left[-\nu_i(\vec{\theta}) + n_i \log \nu_i(\vec{\theta}) + ... \right], \tag{7.6}$$

where N_{bins} is the number of bins in the distribution and the terms not depending on $\vec{\theta}$ are dropped. This is done to simplify the computation, as the absolute value of the likelihood does not hold any information: only the derivation or the relative difference with respect to the maximum have to be known.

7.1.4 Parameter of interest and profile likelihood

The $t\bar{t}H(b\bar{b})$ model contains numerous parameters, the signal strength $\mu_{t\bar{t}H}$ is, however, the most important since it defines the amount of measured signal. Such parameter is generally called a *parameter of interest* (POI). The remaining parameters are then called *nuisance parameters* (NPs), since their value is not necessarily important.

The maximum likelihood procedure ascertains the value of the POI. However, it cannot be directly used to make definitive statements when it comes to e.g. the significance or the limits of the measurement due to an interference of the POI with other parameters. This dependence on the other parameters can be minimized through *profiling*, which removes the dependence on the nuisance parameters in the large sample limit[21].

The likelihood is first redefined as $L(\mu, \vec{\theta})$, where μ is the POI and $\vec{\theta}$ represents the NPs. The *profile likelihood*[21], which depends only on the parameter of interest μ, is defined in the following form:

$$L_P(\mu) = L(\mu, \hat{\hat{\theta}}(\mu)), \text{ where } \frac{\partial L(\mu, \vec{\theta})}{\partial \vec{\theta}}|_{\hat{\hat{\theta}}(\mu)} = 0. \tag{7.7}$$

where one maximizes the likelihood for each value of μ. With a likelihood depending only on μ, one can construct a test statistic:

$$t(\mu) = -2 \log \left(\frac{L_P(\mu)}{L(\hat{\mu}, \hat{\vec{\theta}})} \right) = -2 \log \left(\frac{L_P(\mu)}{L_P(\hat{\mu})} \right), \tag{7.8}$$

which approaches a χ^2 distribution in the limit of large sample size, with $\hat{\mu}$ maximizing the profile likelihood.

Since the $t\bar{t}H(b\bar{b})$ process was not yet observed, the eventual goal of the analysis is its discovery. To simplify the discussion, from this point on the μ represents a signal strength or a signal yield. This means that $\mu = 0$ is the absence of a signal and any positive value means its presence. The true value of μ cannot be negative, though the measured value can.

The statistical significance of the measured μ, assuming it is positive, can be directly determined from the value of the test statistic at $\mu = 0$ [139]:

$$Z_0 = \sqrt{t(0)} = \sqrt{2[\log L_P(\hat{\mu}) - \log L_P(0)]}, \tag{7.9}$$

defined in units of number of standard deviations σ. The value can be also related to the difference between the log-likelihood at μ =0 (or in the absence of a signal) and the measured value $\hat{\mu}$. The ATLAS experiment would report a discovery if $Z_0 \geq 5$, which corresponds to a probability of a false positive of approximately 3×10^{-7}.

7.1.5 Asimov dataset and median significance

In order to estimate the expected significance and other properties based on the prediction of the nominal model, an *Asimov* dataset can be used: a *pseudodata* constructed such that all estimators correspond to their true value[139], but reflecting the statistics of the real data. In the case of a histogram based analysis, as is the $t\bar{t}H(b\bar{b})$, its construction is easy. One simply takes the predicted distribution and assigns each bin an uncertainty based on its yield (corresponding to the Poisson distribution).

The likelihood is constructed as for any other data. The profile likelihood fit performed to such dataset will then, by definition, give the true values for all parameters and it can be shown that the median significance of the model is simply the significance given by the formula 7.9[139], where now the measured value of $\hat{\mu}$ corresponds to the true value of the nominal model. It also provides an approximation of how much the data should be able to constrain various nuisance parameters.

7.1.6 Implementation of additional uncertainties

The approach described in the previous sections only takes into account statistical fluctuations of the data and assumes a precise prediction. However, in reality there are also systematic uncertainties that modify the yield in each bin.

Systematic uncertainties are usually computed for each Monte Carlo sample individually. To simplify the discussion, the following assumes that the prediction ν represents a single

sample and the uncertainties are computed relative to it. A generalization to multiple samples then requires to sum together the modified predictions ν^{new} for each sample.

The analysis works under the assumption that the source of each uncertainty follows a Gaussian distribution. The systematic variations then change the theoretical prediction ν_i to a modified version ν_i^{new}. To implement the systematic variations, a nuisance parameter α_j, following a Gaussian distribution with a mean at 0 and a width set to 1, is defined for each systematic j. Assuming that for each bin i the relative size of the uncertainty is S_i^j, the predicted value gets shifted by $\nu_i(\mu, \vec{\theta}) S_i^j \alpha_j$:

$$\nu_i^{\text{new}}(\mu, \vec{\theta}, \vec{\alpha}) = \nu_i(\mu, \vec{\theta})(1 + \sum_j^{N_{\text{syst}}} S_i^j \alpha_j), \tag{7.10}$$

where N_{syst} is the total number of systematic uncertainties[1].

Additional Gaussian terms $\propto \exp(-\alpha_j^2/2)$ are then added to the likelihood, expressing a penalty on the possible values of the α_j. This expands the log-likelihood in equation 7.6 in the following way:

$$\log L(\mu, \vec{\theta}, \vec{\alpha}) = \sum_i^{N_{\text{bins}}} \left[-\nu_i^{\text{new}}(\mu, \vec{\theta}, \vec{\alpha}) + n_i \log\left(\nu_i^{\text{new}}(\mu, \vec{\theta}, \vec{\alpha})\right)\right] - \sum_j^{N_{\text{syst}}} \alpha_j^2/2, \tag{7.11}$$

where the last term is simply a logarithm of the Gaussian terms apart from constants.

The post-fit value of α_j describes by how many standard deviations the systematic was shifted, or *pulled*, with respect to its default value. The uncertainty on the α_j value then informs how much smaller (or larger) the size of the systematic is after the fit, 1 being the pre-fit value. The uncertainties are usually either not affected or reduced (*constrained*) by the fit.

7.1.7 Pruning of systematics

As summarized later in section 7.2, a large number of systematic uncertainties (over 200) is considered in the analysis. Some of them have a significant impact on the result of the fit, as for example the $t\bar{t}b\bar{b}$ modeling discussed previously, but most of them are negligible. If all of them were included in the fit, the minimization procedure would become very time-consuming. To avoid this, systematics which would have a negligible effect are omitted. This is done independently for each analysis region.

First, systematic variations are split into their normalization and shape components[2]. The former is dropped, if the normalization difference with respect to the nominal distribution is smaller than 0.5%. Similarly, a shape of a systematic is dropped when no bin of the shape component in the given region has a relative difference bigger than 0.5%. In order to assure that the choice of the pruning threshold does not have a significant effect on the fit result, a lower cut-off value of 0.1% for both the shape and the normalization component was tested, and no notable difference was found.

[1] In theory, one could combine the $S_i^j \alpha_j$ factors and have a Gaussian with corresponding width, but the separation allows for a better interpretation of the fit results: $\alpha_j = 0$ corresponds to the nominal value, while $\alpha_j = \pm 1$ means that the parameter was shifted by ± 1 of its pre-fit uncertainty.

[2] Normalization component simply describes effect on the yield in a given region. Shape component is then the systematic modified such that it does not affect the normalization in given region.

7.1.8 Goodness of fit

The goodness of fit in the analysis is evaluated using a *saturated model*. The saturated model has one extra free parameter per bin of the distribution, thus being able to perfectly fit the data. The ratio between the likelihood of the nominal model and of the saturated model then follows a χ^2 distribution asymptotically[140] and the corresponding χ^2 probability[138] is taken as a goodness of fit value.

7.1.9 Software implementation

The analysis relies on an internal ATLAS fitting framework called TRExFitter, which builds histograms from the input data and provides them to the tools for statistical analysis. The fit itself is done using the HistFactory package[141], a tool specifically designed for profile likelihood fits in form of histograms. HistFactory is built on the RooFit[142] and RooStats[143] packages, which provide necessary tools for the performance of the fit.

The minimization of the likelihood is done using the Minuit algorithm[144] implemented in ROOT, a C++ based framework for data analysis [133]. Uncertainties on the parameters are by default described using the quadratic approximation, while for more interesting parameters the more precise graphical method is used. Both methods were described previously in section 7.1.2.

7.2 Summary of the nominal statistical model

The nominal statistical model of the $t\bar{t}H(b\bar{b})$ analysis, used in the profile likelihood fit, was described in chapter 6. It contains two free-floating parameters with no prior uncertainty: the signal strength $\mu_{t\bar{t}H}$, which represents normalization of the signal, and $k(t\bar{t}+\geq 1b)$, the normalization of the $t\bar{t}+\geq 1b$ component.

Among the systematic uncertainties, two groups are the most important: the $t\bar{t}H$ and $t\bar{t}+\geq 1b$ modeling uncertainties, which are divided into the following four groups: ISR, FSR, PS&had and NLO match, as described in sections 6.8.1 and 6.8.2 respectively.

All systematic uncertainties are summarized in table 7.1, which lists the main systematic categories and the number of their components. Furthermore, whether the systematic uncertainty affects only the normalization or also the shape is specified.

The actual list of systematic variations used in the fit is smaller due to the pruning procedure described in section 7.1.7. Most of the uncertainties have a negligible impact and are removed to improve the convergence of the fit.

The combined statistical uncertainty of the nominal Monte Carlo samples is added as an additional systematic uncertainty for each bin of the analyzed distribution.

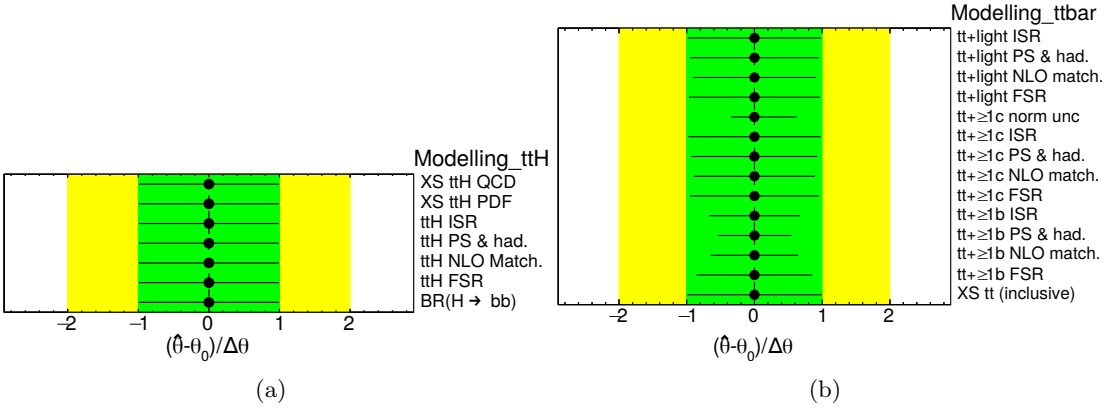
7.3 Fit to the Asimov pseudodata

The first test performed with the complete $t\bar{t}H(b\bar{b})$ model is a fit to the Asimov dataset. This is done to estimate the significance and to see how the expectation will get constrained by the fit. Furthermore, it determines which systematics will have a sizable effect on the sensitivity.

The fit results in a signal strength $\mu_{t\bar{t}H} = 1.00^{+0.53}_{-0.49}$, corresponding to a median significance 2.0σ, derived as described in section 7.1.4. The free floating normalization of the $t\bar{t}+\geq 1b$ is $k(t\bar{t}+ \geq 1b) = 1.00^{+0.09}_{-0.08}$. The contribution of the data statistical uncertainties to the result is estimated by repeating the fit with all systematic parameters

Systematic uncertainty	Type	Comp.
Experimental uncertainties		
Luminosity	N	1
Pileup modeling	SN	1
Physics Objects		
Electrons	SN	7
Muons	SN	15
Jet energy scale	SN	31
Jet energy resolution	SN	9
Jet vertex tagger	SN	1
$E_{\mathrm{T}}^{\mathrm{miss}}$	SN	3
b**-tagging**		
Efficiency	SN	45
Mis-tag rate (c)	SN	20
Mis-tag rate (light)	SN	20
Signal and background modeling		
Signal		
$t\bar{t}H$ cross-section	N	2
H branching fractions	N	3
$t\bar{t}H$ modeling	SN	4
$t\bar{t}$**+jets Background**		
$t\bar{t}$ cross-section	N	1
$t\bar{t}+\geq 1c$ normalisation	N	1
$t\bar{t}+\geq 1b$ normalisation	N (free floating)	1
$t\bar{t}$+light modeling	SN	4
$t\bar{t}+\geq 1c$ modeling	SN	4
$t\bar{t}+\geq 1b$ modeling	SN	4
Other Backgrounds		
cross-section	N	18
modeling	SN	9

Table 7.1: List of systematic uncertainties included in the analysis. An "N" means that the uncertainty is taken as normalisation-only for all processes and channels affected, whereas "SN" means that the uncertainty is taken on both the shapes and the normalisation. Some of the systematic uncertainties are split into several components for a more accurate treatment: the number of such components is indicated in the column labeled as "Comp.". Courtesy of the $t\bar{t}H(b\bar{b})$ analysis team.



(a) (b)

Figure 7.2: Resulting pulls and constraints of the $t\bar{t}H$ (a) and $t\bar{t}$+jets (b) uncertainties for fit to the Asimov dataset.

except for the free-floating parameters $\mu_{t\bar{t}H}$ and $k(t\bar{t}+\geq 1b)$ fixed to their estimated value (in the case of the Asimov fit this mean the true values of all parameters). The $t\bar{t}+\geq 1b$ normalization is basically unaffected by the statistics with only 2% uncertainty. For the $\mu_{t\bar{t}H}$ the effect is relatively large (22%), but still much smaller than the overall uncertainty: the analysis is limited by the systematic uncertainties.

The $t\bar{t}$+jets and the $t\bar{t}H$ modeling systematic uncertainties and their constraints are displayed in figure 7.2. The modeling systematics of the $t\bar{t}H$ are not constrained at all, indicating low sensitivity to the $t\bar{t}H$ modeling with the given precision.

Since there is only a small $t\bar{t}$+light contribution in the analysis regions, its systematic variations also remain unconstrained. For the $t\bar{t}+\geq 1c$ there is a slight effect on the shape through the modeling uncertainties, but more importantly there is a significant constraint of the $t\bar{t}+\geq 1c$ normalization uncertainty. The most crucial aspect is the $t\bar{t}+\geq 1b$ modeling, where the constraints are large. That is especially the case of the two-point systematic variations: the PS&had and the NLO match.

The likelihood fit also provides correlations between the various parameters of the model, based on the variance shown in formula 7.3. Large correlations can point to unnecessary degrees of freedom which are covered by other parameters, or to some variables being too correlated to the signal, making the analysis less sensitive. Those parameters can be further investigated to find out if their impact can be reduced.

The correlation matrix for the Asimov fit can be found in figure 7.3. The $t\bar{t}+\geq 1b$ normalization is correlated to the $t\bar{t}+\geq 1b$ ISR and PS systematics. This correlation will be reduced in the combination with the dilepton channel, because the systematics are normalized such that they do not change normalization in the combined phase-space (as discussed in section 6.7.4). The signal is strongly correlated to the $t\bar{t}+\geq 1b$ NLO match systematic. This implies that the systematic has similar features as the signal, which can lead to a lower sensitivity. To estimate the actual effect of the systematics on $\mu_{t\bar{t}H}$, a *ranking* plot is constructed, which ranks nuisance parameters based on their impact on the signal. This impact can be derived by fixing the nuisance parameter α_i of the given systematic to:

- ± 1 (the pre-fit value) for the **pre-fit impact**
- its post-fit value for the **post-fit impact**

and by repeating the fit. The displayed impact is then the difference in $\mu_{t\bar{t}H}$ between the nominal fit and the variation.

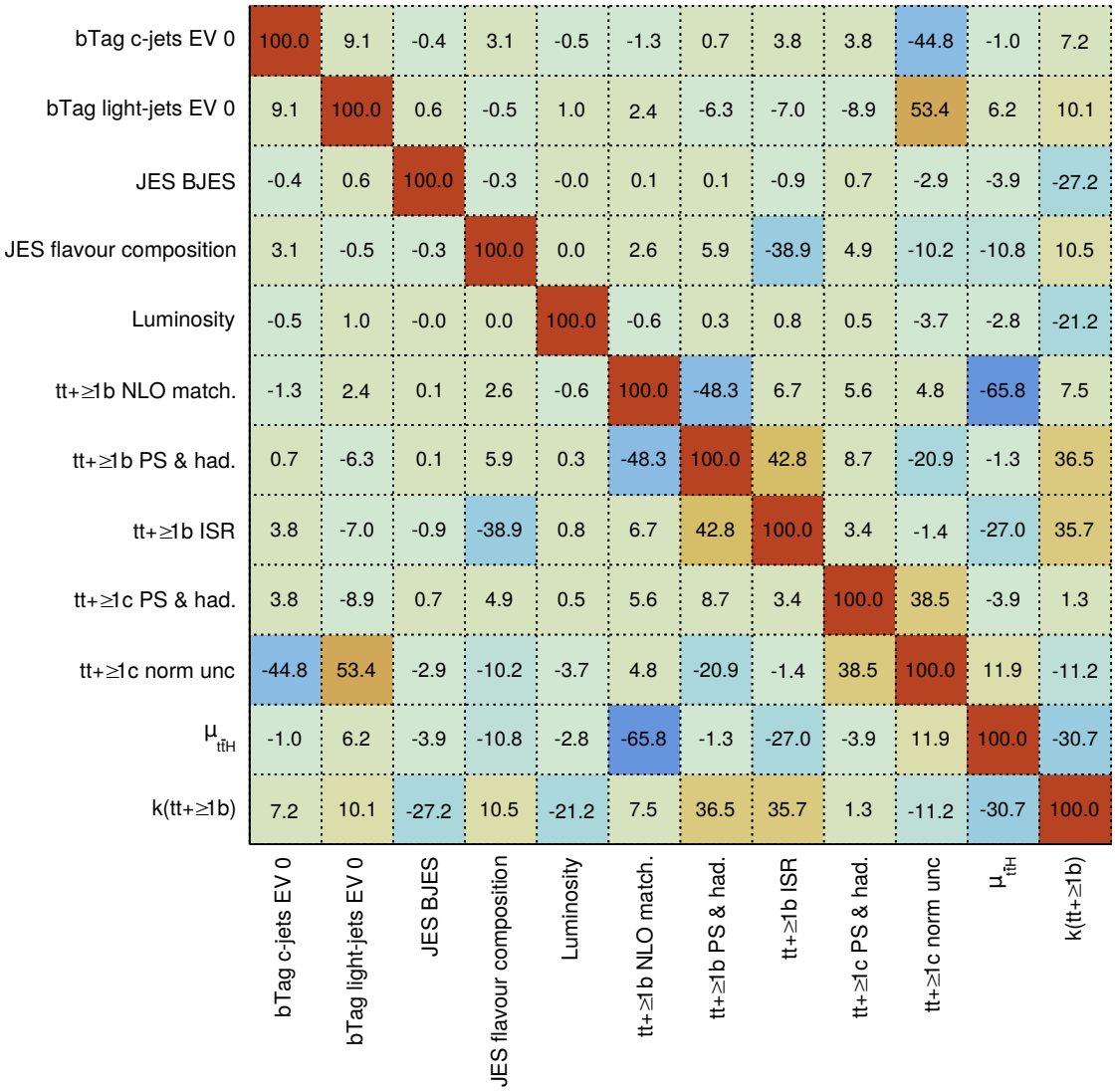


Figure 7.3: Correlation matrix for the fit to the Asimov dataset, showing all parameters which have at least one correlation larger than 20%.

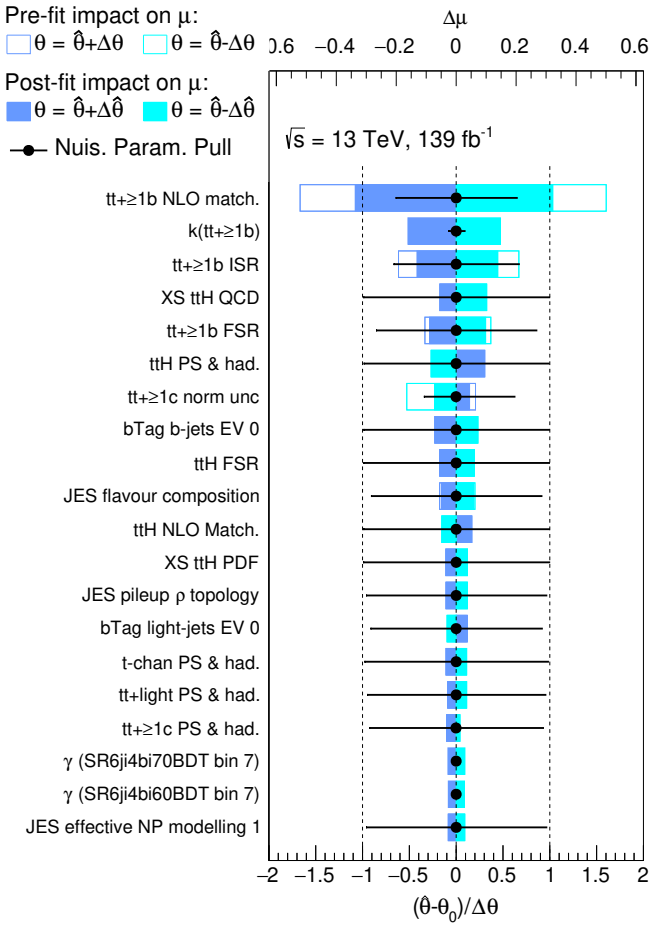


Figure 7.4: Ranking plot for the 20 nuisance parameters with the largest impact on the parameter of interest $\mu_{t\bar{t}H}$. It is shown for the pre-fit uncertainty by the empty box and post-fit by the filled box. The post-fit shifts and constraints of the systematics is displayed by the black markers and the horizontal line respectively. Vertical dashed lines correspond to ± 1 values of the nuisance parameters.

For the Asimov fit the ranking plot is displayed in figure 7.4. The nuisance parameters are ordered by the size of their post-fit effect on $\mu_{t\bar{t}H}$. The first 20 are shown, though only a handful have a large impact on the result. In addition, the plot also shows the constraints and pulls of the systematics variations, similarly to the plots in figure 7.2.

The impact on $\mu_{t\bar{t}H}$ for both the pre-fit and post-fit is dominated by the $t\bar{t}+\geq 1b$ NLO match systematic uncertainty. By shifting the systematic by 1σ of its post-fit uncertainty, the $\mu_{t\bar{t}H}$ is shifted by around 30%.

The $t\bar{t}+\geq 1b$ NLO match uncertainty has approximately twice the impact of the next two systematics with largest impact: the $k(t\bar{t}+\geq 1b)$ normalization and the $t\bar{t}+\geq 1b$ ISR.

The effect of the next few systematic variations, coming from $t\bar{t}H$ and $t\bar{t}+\geq 1b$ modeling, is relatively similar, though with a slowly decreasing effect. The first experimental systematic uncertainty in the ranking is the first eigenvalue of the b-tagging efficiency of b-jets (bTag b-jets EV0) at the eighth place, confirming the low impact of experiemental uncertainties on the result of the measurement.

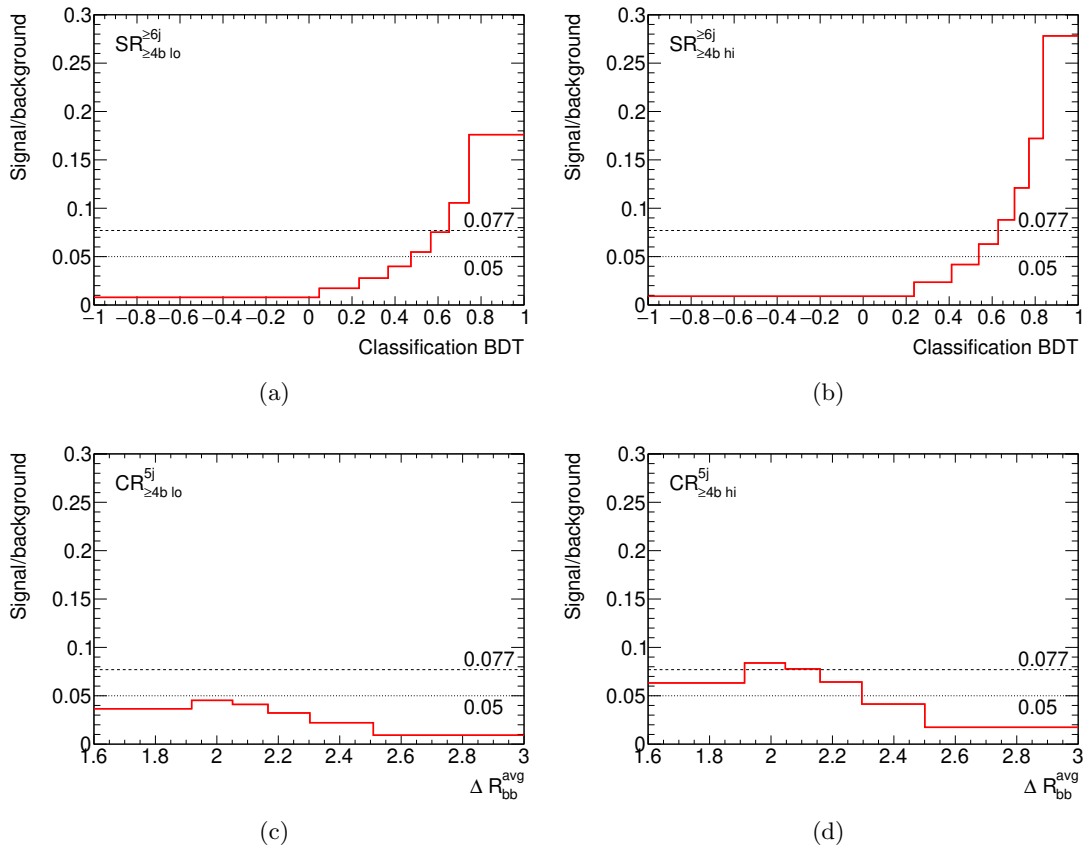


Figure 7.5: The signal over background composition of the four analysis regions. The horizontal lines display the 5% and 7.7% threshold used for the blinding in the analysis, where the bins with S/B above the lines are removed from the fit.

7.4 Blinding strategy

Fits to the Asimov pseudodata give the expected performance of the nominal model. However, especially in the case of an analysis with a large mis-modeling, the real sensitivity and performance will differ significantly in the fit to the data.

Simply performing the nominal fit to the data without excluding the signal and optimizing the model based on its results could lead to a bias, whether conscious or unconscious, modifying the model to acquire a desired result. In order to avoid this, a blinding procedure is implemented: bins with a large expected contribution of the signal are removed from the fit and the data are not displayed in the plots. This is an intermediate step made to validate the nominal model.

In this thesis most of the results prior to the unblinding will be with a 7.7% blinding threshold on the signal over background ratio[3]. Distributions of the signal over background ratio in the four analysis regions can be found in figure 7.5. The following will therefore distinguish between fits in the *full* range and with the *blinded* data with a limited phase-space.

[3]The blinding threshold used to be 5% of the expected signal over background ratio, providing regions with a low sensitivity to the signal. This was later increased to 7.7%, which corresponds to 1 extra unblinded bin per region on average. This was done in order to test if the results of the tests done in the revealed bins does not change in the few extra bins before unblinding fully. This proved be the case.

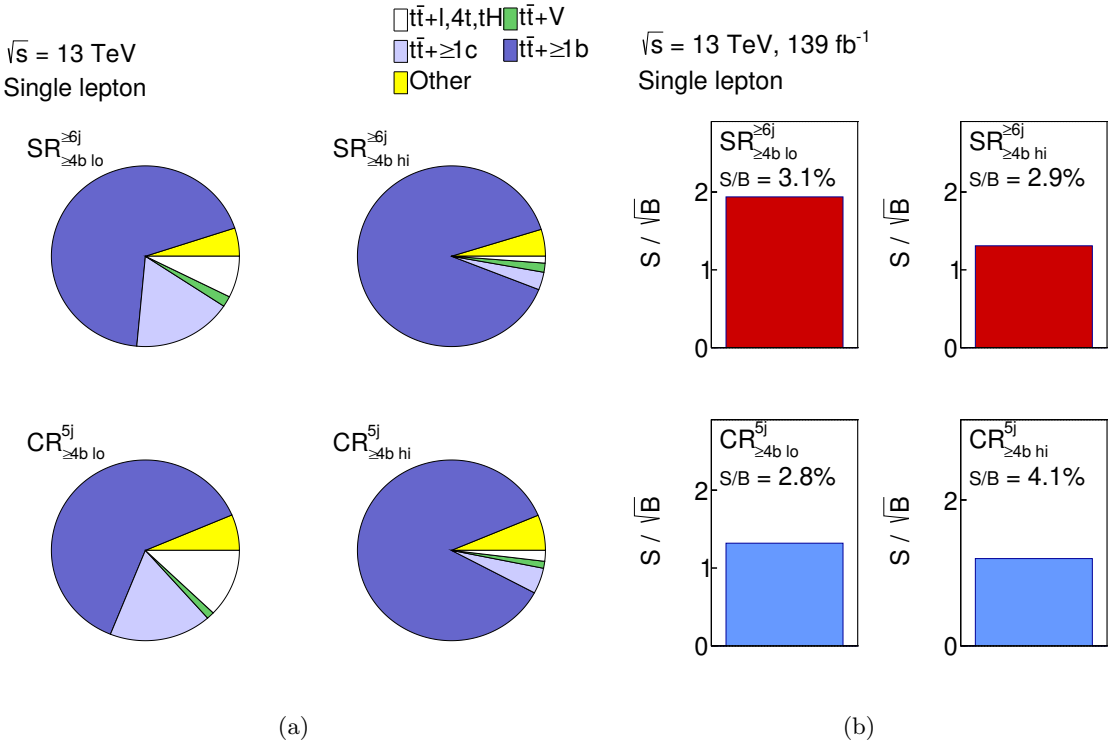


Figure 7.6: Expected background composition (a) and the expected signal over background ratio and the statistical significance (b) in the four analysis regions in the revealed bins.

The expected background composition of the revealed bins under the 7.7% threshold can be found in figure 7.6(a), which show similar fractions of background as the full model (shown previously in figure 6.4(a)). The amount of signal is displayed in figure 7.6(b), giving around 3-4% on average per region.

7.5 Background-only fits to blinded data

Background-only fits, or fits omitting the $t\bar{t}H$ signal and its modeling uncertainties in the likelihood, are performed in the revealed (not blinded) bins in order to test the background model and to get a better idea about its expected performance. This comes with several caveats: the distributions of the model can be similar in the revealed bins but differ in the blinded bins, leading to different correlation and thus different post-fit values.

The background-only fit also completely ignores the signal. Even though its contribution in the revealed bins is low (as was shown in the previous section), its absence in the model can still slightly bias the fit. These effects are explored later in the context of the pseudodata fits.

Since the signal is omitted, $k(t\bar{t}+\geq 1b)$ is the only free-floating factor. Its post-fit value is found to be $1.16^{+0.10}_{-0.09}$, with 2% statistical uncertainty. The nuisance parameters are now shifted with respect to their nominal value, as can be seen in figure 7.7.

Experimental systematics are only slightly pulled and constrained as are the systematics of the non-$t\bar{t}$+jets backgrounds. For the $t\bar{t}$+jets systematics both the pulls and the constraints are larger. Most of them are still compatible with the nominal value 0 within

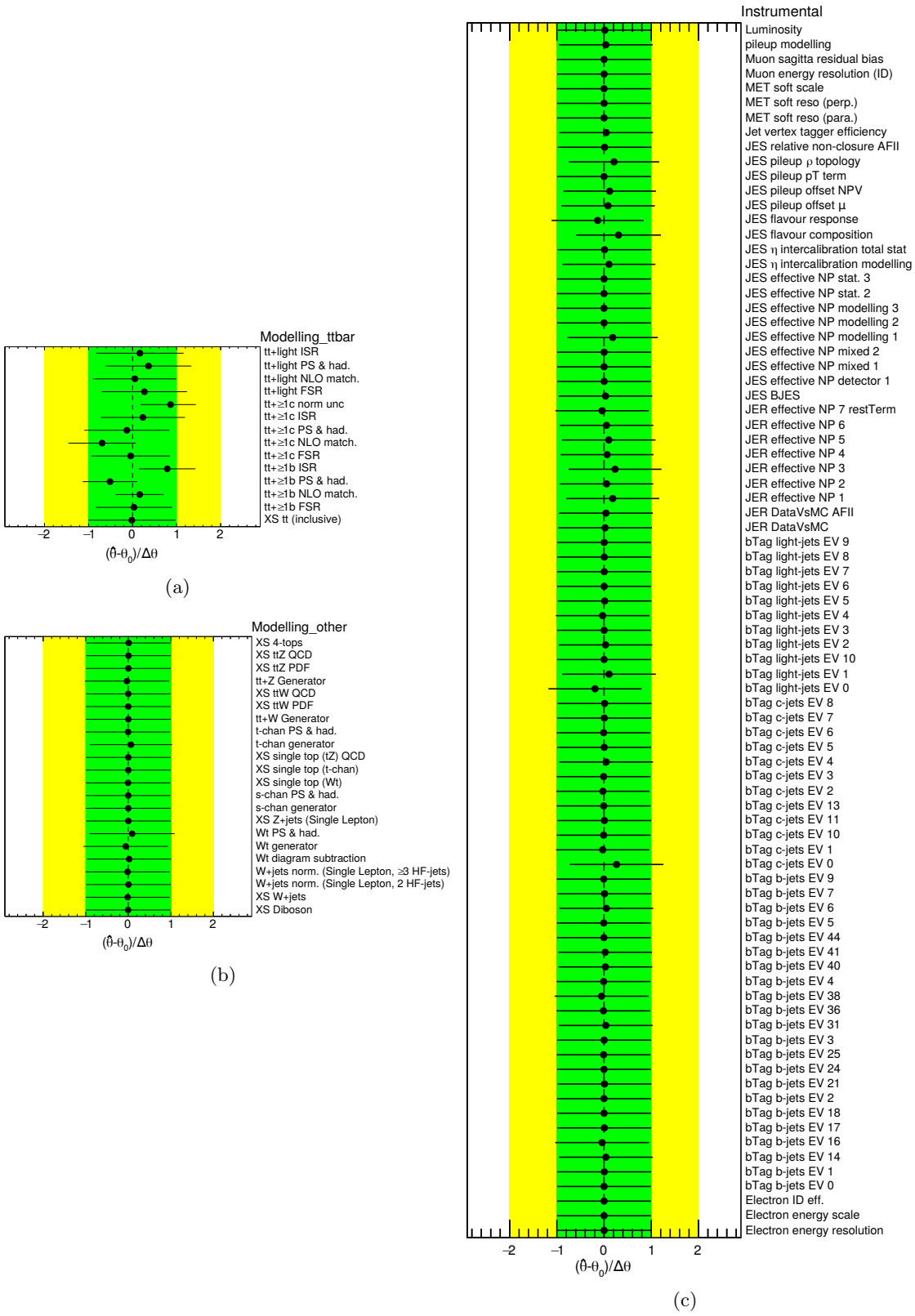


Figure 7.7: Resulting pulls and constraints of the $t\bar{t}$+jets modeling (a), modeling of other backgrounds (b) and instrumental (experimental) systematic uncertainties (c) for the background-only fit in the revealed bins.

one standard deviation. There are only two exceptions where the pulls are large: the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalization, and the $t\bar{t}+\geq 1b$ ISR systematic. The latter mainly affects the n_{jets} distribution, as is further explored in the next chapter in section 8.3.1. The measured values of the $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ normalizations (approx. 1.2±0.1 and 1.8±0.6) are in agreement with the previous ATLAS measurement of the $t\bar{t}H(b\bar{b})$ (1.24 and 1.63)[8] and an independent ATLAS measurement of the $t\bar{t}b\bar{b}$ (1.1 and 1.6)[122]. The estimation of the normalization will be more precise in the combination with the dilepton channel, where the contribution of the $t\bar{t}+\geq 1c$ is larger.

Post-fit distributions of the analysis regions can be found in figure 7.8. They show good agreement, with χ^2 probability[4] larger than 90% in most of the regions. Only the $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ region shows a a slightly lower probability due to a single bin further from the fitted distribution, most probably just due to a statistical fluctuation, especially since a similar feature is not present in the tighter $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ region.

7.6 Data-driven expectation

The Asimov fit naively assumes correctness of the nominal background model, and as such the significance derived through it may not reflect the result of the fit to the data in the full phase-space, especially given the large mis-modeling. However, the background-only fit provides a good first approximation of the background modeling, which can be used to derive a more realistic estimation of the expected significance.

One way to do such estimation is to propagate the post-fit values of the model parameters, derived in the fit to the blinded data, to the full range. This can be done by constructing pseudodata from the nominal model, similarly to Asimov pseudodata, but shifting values of the model parameters to correspond to the background-only fit to the blinded data.

Since the signal was not included in this fit, it is added to these pseudodata unchanged. The pseudodata constructed this way will be referred to as *data-driven pseudodata*, a simple extrapolation of the measured values in the revealed bins to the inclusive distribution.

This model carries some caveats. The first caveat is the absence of the signal in the background-only fit. To estimate the impact of this, the background-only fit is repeated with a fixed amount of signal, specifically $\mu_{t\bar{t}H} = 1$ and $\mu_{t\bar{t}H} = 2$ and additional pseudodata are constructed from the post-fit values of these fits. This way, three pseudodata are created, designated with additional label $\mu^{fixed}_{t\bar{t}H} = 0$, $\mu^{fixed}_{t\bar{t}H} = 1$ and $\mu^{fixed}_{t\bar{t}H} = 2$ which refers to the type of fit they are derived from. The pseudodata itself, however, contain the nominal amount of signal ($\mu_{t\bar{t}H} = 1$), as this value is expected in the full range.

Another caveat of this approach is the difference in correlations between the full range and the blinded range, which will lead to slightly different behavior of the background model. Nevertheless, the data-driven pseudodata still provide a better estimation of parameters than a simple fit to the Asimov pseudodata.

The signal strengths of the three fits, based on pseudodata from the $\mu^{fixed}_{t\bar{t}H} = 0$, $\mu^{fixed}_{t\bar{t}H} = 1$ and $\mu^{fixed}_{t\bar{t}H} = 2$ background-only fit, are shown in figure 7.9. They have a slight dependency on the type of the pseudodata (approx. 17%), as does the normalization of $t\bar{t}+\geq 1b$ as displayed in figure 7.10, though there the shift is only 6%.

Finally, nuisance parameters of the $t\bar{t}$+jets background, shown in figure 7.11(a), are generally similar, showing a small dependence on the amount of signal in the revealed bins. The biggest change is for $t\bar{t}+\geq 1b$ NLO match.

[4]The χ^2 probability is computed with both statistical and systematic components of the uncertainty[138].

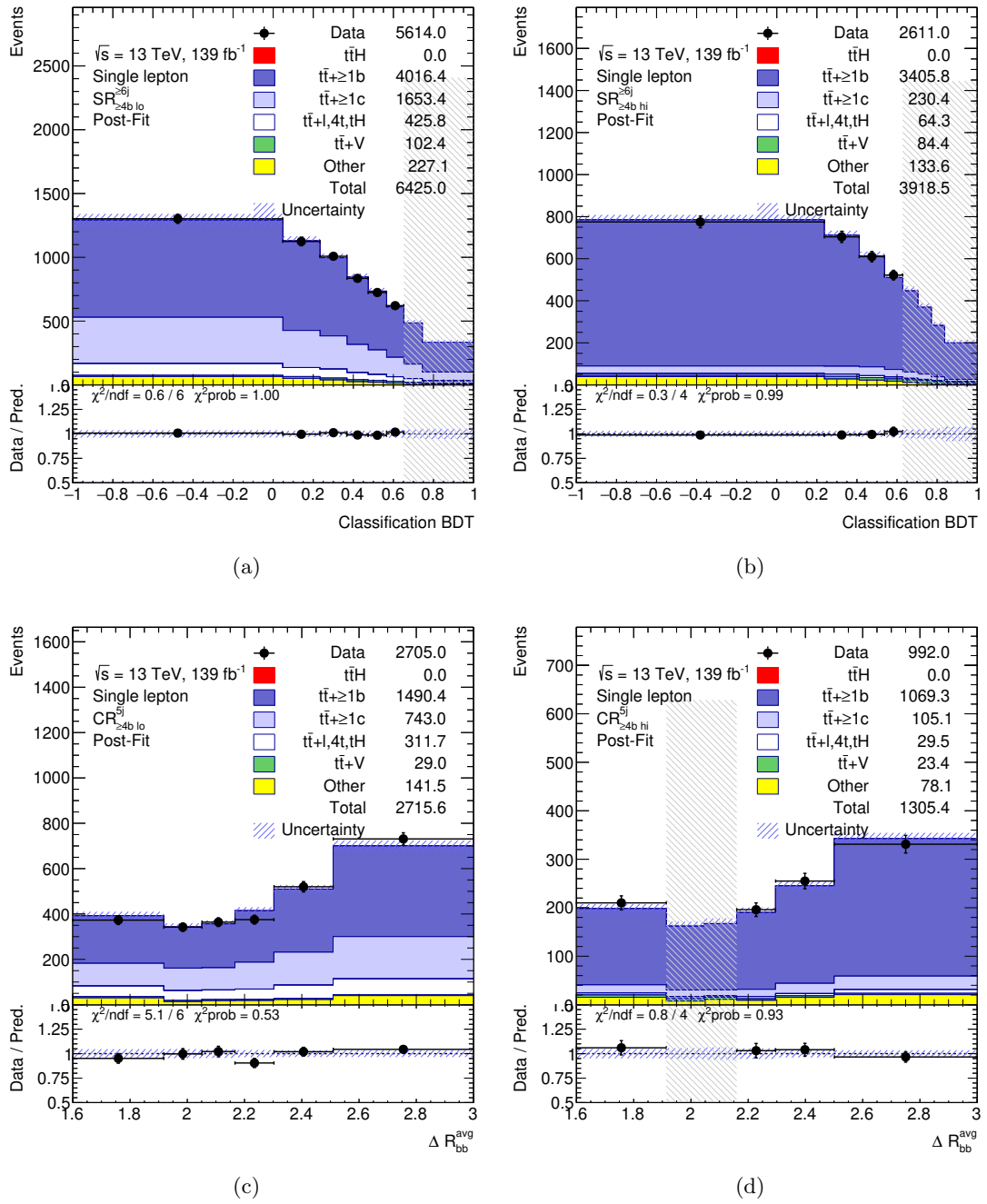


Figure 7.8: The post-fit distributions of the Monte Carlo compared to the data in the four analysis regions, (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$, (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$, (c) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{lo}}$ and (d) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{hi}}$, after performing the background-only fit, displayed as a function of the discriminating variable used: $\Delta R^{\mathrm{avg}}_{bb}$ for the 5jet regions and the classification BDT for the 6-jet regions. The gray shaded bands designate the blinded bins.

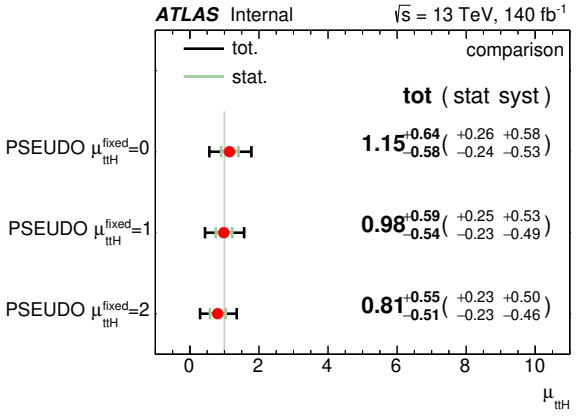


Figure 7.9: Resulting values and uncertainties of the $\mu_{t\bar{t}H}$ normalization factor, comparing data-driven pseudodata created based on a background-only fit with three values of $\mu_{t\bar{t}H}^{fixed}$ (0, 1 and 2), which has to be distinguished from $\mu_{t\bar{t}H}$ derived from the subsequent fit to the psedudodata.

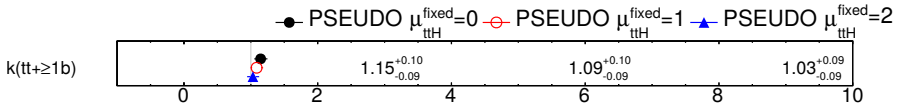


Figure 7.10: Resulting values and uncertainties of the $t\bar{t}+\geq 1b$ normalization factor, comparing data-driven pseudodata created based on a background-only fit with three values of $\mu_{t\bar{t}H}^{fixed}$ (0, 1 and 2).

Parameters of the $t\bar{t}H$ modeling, displayed in figure 7.11(b), are centered at 0 and correspond to the model used to generate the pseudodata.

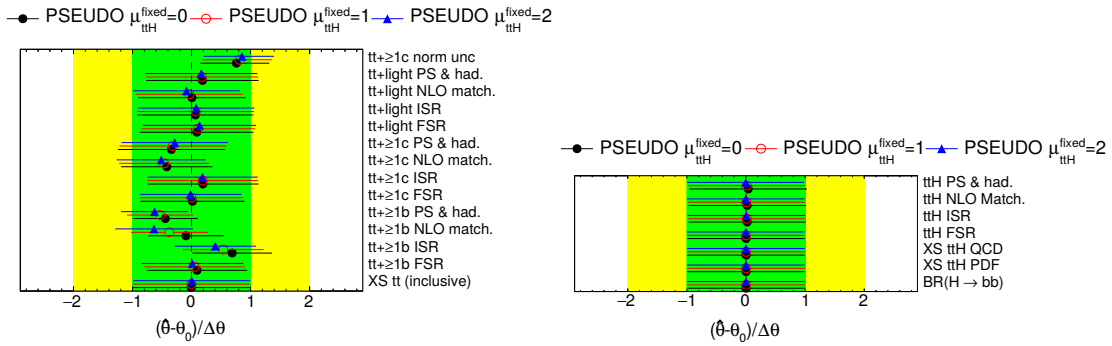
The fit with $\mu_{t\bar{t}H}^{fixed}$=0 gives a median significance of 1.8σ, only marginally smaller than the expected value from the Asimov fit (2.0σ). This is a simple consequence of the larger background measured in the data, since both the $t\bar{t}+\geq 1b$ and the $t\bar{t}+\geq 1c$ have larger normalization with respect to the nominal prediction. In comparison, the fit with $\mu_{t\bar{t}H}^{fixed}$=2 gives a significance of 1.9, meaning that the amount of the signal in the revealed bins does not have a large impact on the estimation.

To summarize, in the full fit to data a slightly lower significance and a small bias on the $\mu_{t\bar{t}H}$ can be expected, based on the performance of the background-only fit and its extrapolation to the full phase-space. This bias is reduced in the combination with the dilepton channel as discussed in the next chapter.

7.7 Pseudodata based on alternative models

One way to estimate the robustness of a model is to create pseudodata with different generators and performing the various types of fits mentioned in previous sections. For the $t\bar{t}H(b\bar{b})$ process, testing the $t\bar{t}b\bar{b}$ modeling is the most important so alternative $t\bar{t}+\geq 1b$ models should be tested.

Two Monte Carlo samples are available: Sherpa $t\bar{t}b\bar{b}$, which is the most similar to the nominal as it has the $t\bar{t}b\bar{b}$ generated as part of the matrix element, and the inclusive $t\bar{t}$+jets PowhegBox+Pythia8 generator, which has significantly different modeling (mainly b quarks coming from the parton shower). Both samples were described previously in section 6.4.3.



(a) (b)

Figure 7.11: Resulting pulls and constraints of the $t\bar{t}H$ (b) and $t\bar{t}$+jets (a) uncertainties, comparing data-driven pseudodata created based on a background-only fit with three values of $\mu_{t\bar{t}H}^{fixed}$ (0, 1 and 2).

In comparison to the data-driven pseudodata shown in the previous section, this approach has several advantages. The modeling is known in the full phase-space before unblinding, so one can test how the fit is able to correct the $t\bar{t}+\geq 1b$ modeling or see how it affects the performance of other backgrounds or the signal in the fit. Furthermore, one can compare the nominal fit in the full phase-space to the background-only fit in revealed bins, and the data-driven pseudo-dataset extracted from it.

These three types of fits are presented for the SHERPA pseudodata, where the data-driven extrapolation is studied only for the extrapolation from $\mu_{t\bar{t}H}$ =0. Resulting values of the $t\bar{t}+\geq 1b$ normalization and $t\bar{t}$+jets modeling systematics can be found in figures 7.12(b) and 7.12(c) respectively, where the lower $t\bar{t}+\geq 1b$ normalization is the result of a different cross-section of the SHERPA generator. The post-fit values of the background modeling uncertainties differ only slightly between the three fits, with the largest difference in the $t\bar{t}+\geq 1b$ NLO match modeling systematic. That is not surprising, since the systematic has a larger impact in the blinded bins, which will drive more its post-fit value.

The value of the POI is shown in figure 7.12(a). The results differ by approximately 20% between the full fit to the data and the fit to data-driven pseudodata, a difference smaller than 0.5σ of their uncertainties. Both are compatible with the Standard Model value of $\mu_{t\bar{t}H} = 1$.

The same fits were also done for the $t\bar{t}$+jets sample replacing the $t\bar{t}+\geq 1b$ component. The $t\bar{t}+\geq 1b$ normalization and $t\bar{t}$+jets systematic variations are shown in figures 7.13(b) and 7.13(c), respectively. They have similar features as the previous fit: only the $t\bar{t}+\geq 1b$ NLO match systematic varies significantly between the different types of fit. The measured value of $\mu_{t\bar{t}H}$, shown in figure 7.13(a), though covered by uncertainties is still 35% lower than one, or by approximately $2/3\sigma$. This is the result of the fact that the $t\bar{t}$+jets sample is less signal-like than the nominal $t\bar{t}b\bar{b}$ sample and part of the signal is used to compensate for the difference.

The result also shows a small effect on the $t\bar{t}+\geq 1c$ modeling in the resulting fit, even though the underlying $t\bar{t}+\geq 1c$ model did not change. This suggests that the parameters of the $t\bar{t}+\geq 1c$ modeling are not well controlled in the single lepton fit, especially the $t\bar{t}+\geq 1c$ normalization which changed by almost 20%. This is one of the factors which improves in the combination with the dilepton channel, which has regions with a higher $t\bar{t}+\geq 1c$ contribution. Finally, the $t\bar{t}+\geq 1b$ NLO changes between the background-only and full fit, similarly as for the Sherpa pseudodata fit.

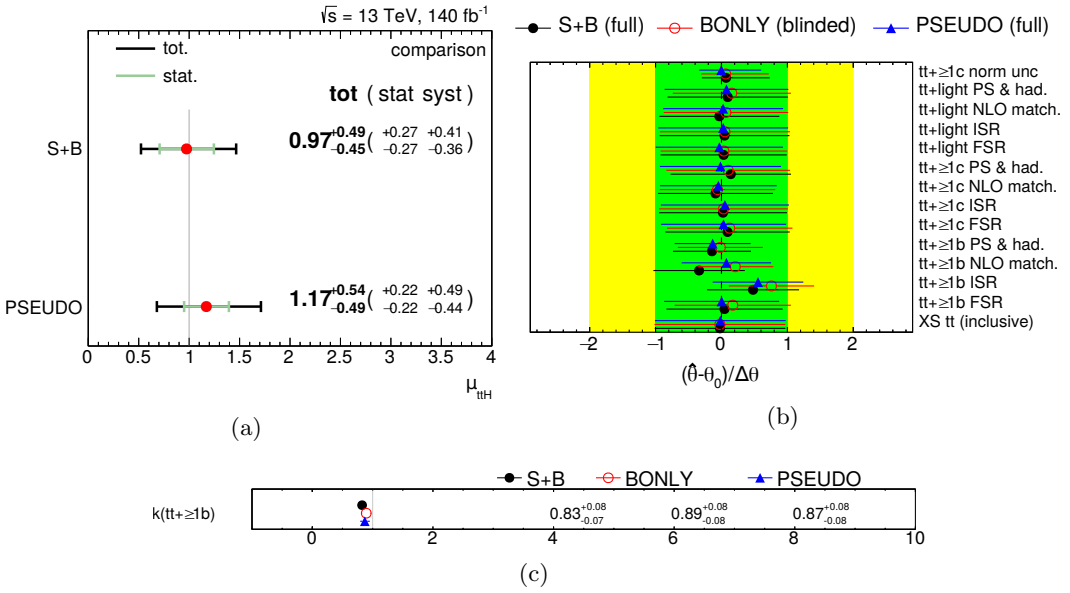


Figure 7.12: Summarizing plots of the three fits to the Sherpa pseudodata: the nominal fit to the data (S+B), background-only fit in the revealed bins (BONLY) and a fit to the data-driven pseudodata based on the results of the background-only fit (PSEUDO). (a) shows values of the parameter of interest $\mu_{t\bar{t}H}$, (c) of the $t\bar{t}+\geq 1b$ normalization $k(t\bar{t}+\geq 1b)$ and finally (b) pulls and constraints of the $t\bar{t}$+jets modeling.

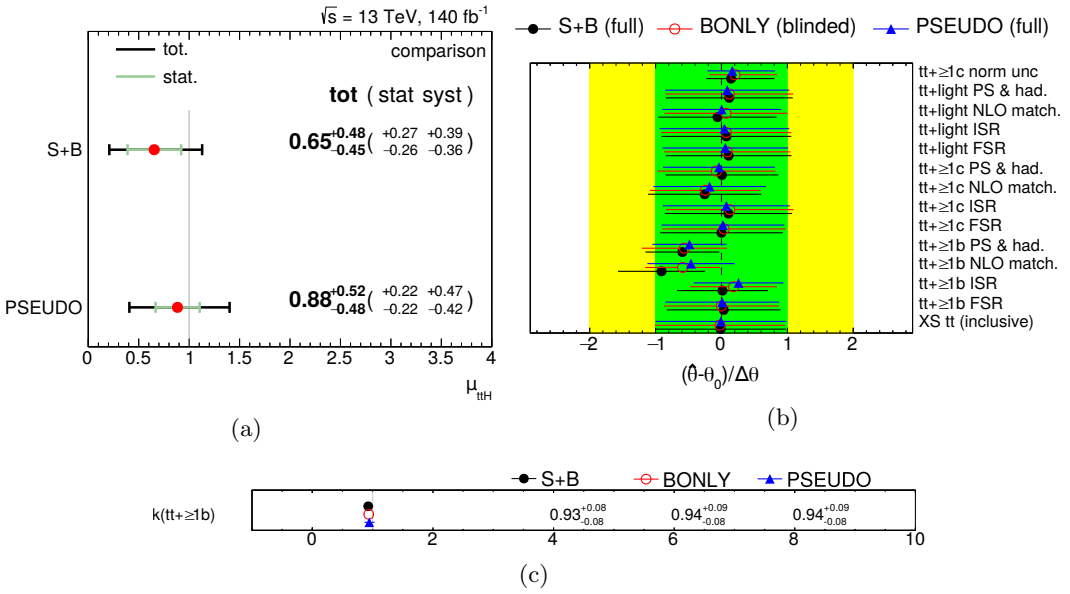


Figure 7.13: Summarizing plots of the three fits to the Powheg +Pythia8 $t\bar{t}$+jets pseudodata: the nominal fit to the data (S+B), background-only fit in the revealed bins (BONLY) and a fit to the data-driven pseudodata based on the results of the background-only fit (PSEUDO). (a) shows values of the parameter of interest $\mu_{t\bar{t}H}$, (c) of the $t\bar{t}+\geq 1b$ normalization $k(t\bar{t}+\geq 1b)$ and finally (b) pulls and constraints of the $t\bar{t}$+jets modeling.

Beside the lower value of the $\mu_{t\bar{t}H}$ for the second pseudodata, another general take-away from these tests is that the background-only fit does not constrain the value of the $t\bar{t}+\geq 1b$ NLO match systematic, since its effect comes mainly from the most signal-like bins. Also the $t\bar{t}+\geq 1c$ has a degeneracy in the single lepton regions because it contributes only slightly to the overall yield. Both effects seem to be partially mitigated in the combination with the dilepton channel, as discussed in chapter 8.

7.8 Impact of statistical fluctuations in the Monte Carlo

For the statistical nominal model, the combined impact of Monte Carlo fluctuations used for the nominal prediction is included in the fit as an additional systematic. This is not the case for the fluctuations of the alternative predictions used to derive the systematic uncertainties, which are smoothed instead to mitigate such effects. While the smoothing should reduce the impact of the statistical fluctuations, it will not completely erase it. This was especially an issue of the previous iteration of the analysis, where the statistics of the Monte Carlo samples was one of the main uncertainties on the measurement (see section 6.1). It is therefore important to study the effect, and to show that the smoothing does not bias the result.

If histograms used in the fit were simply filled with events without any weight, estimating the statistical effects would be simple. Since the content of the histograms follows a Poisson distribution, one can create a toy dataset by varying each bin based on a Poisson distribution with a mean value equal to the bin content. One can easily produce a large number of such datasets and study how the values measured in the fit change.

However, in the case of the Monte Carlo samples used in the $t\bar{t}H(b\bar{b})$ analysis various event weights are used to compensate for detector effects or to increase the precision of the simulation. Their distribution is a priori unknown and their relatively large number would make it difficult to evaluate the statistical effect directly. Hence, it is easiest to treat the input as an unknown distribution.

7.8.1 Bootstrap method

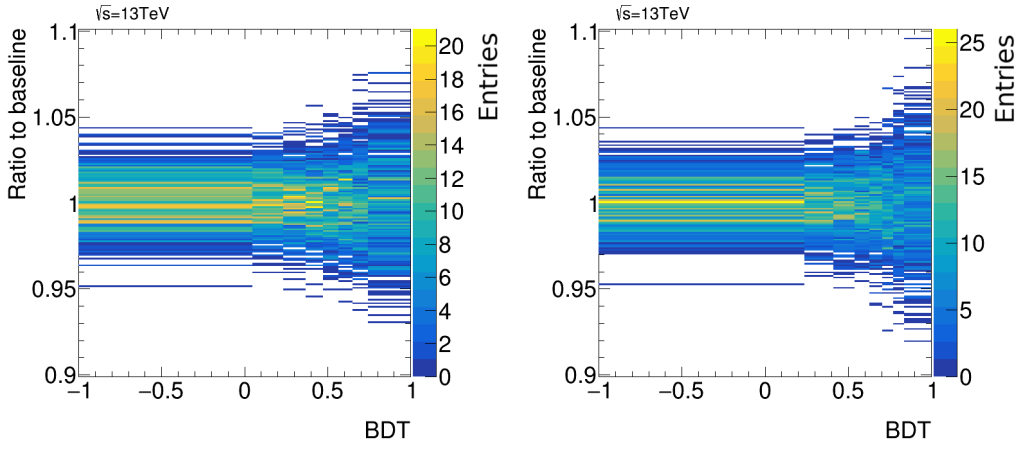
When estimating statistical effects of an unknown distribution, a well established tool is the bootstrap method[145]. Its advantage is that it assumes complete ignorance of the underlying distribution. For n observed events, one first constructs a probability distribution function, a *bootstrap PDF*, by assigning a probability $1/n$ on each event (or $w_i/\sum_j w_j$ for weighted distribution, assuming w_i is the weight of an event i). Then, a *bootstrap sample* is constructed by randomly sampling events from the bootstrap PDF. This means that some events will be used multiple times while others will not enter the histogram at all. The total number of events depends on the context, it can be either the same as the original distribution, or it can be for example smeared by a Poisson distribution.

One of the important properties of the nominal $t\bar{t}H(b\bar{b})$ model is that some systematic variations are correlated to the nominal sample, e.g. by using the same events with different event weights. This is for example the case of the $t\bar{t}+\geq 1b$ ISR and FSR systematic uncertainties. Producing bootstrap samples for both the nominal and the systematic variation independently would lead to an inflation of the statistical fluctuations. To avoid this, the construction of the bootstrap samples is redefined using event *bootstrap weights*, where equivalent weights are used to smear both the nominal sample and the systematic variation simultaneously. The bootstrap weight b_i is then sampled from a

Poisson distribution with a mean value of 1, which produces the same type of bootstrap sample as the PDF construction by weighting each event by $w_i b_i$. Then for a correlated sample, e.g. sample with an alternative event weight w_i', one uses the same bootstrap weight, resulting in a new weight $w_i' b_i$. Beside accounting for the correlated samples, this approach is also practically easier to implement.

In the following sections, the impact of the fluctuations is studied by producing the boostrap samples for a Monte Carlo samples of interest and repeating the fit. The impact on various measured properties, e.g. signal strength or pull and constraint of a systematic uncertainty, is then studied. The statistical impact is then expressed in standard deviations of the measured statistical fluctuations in units of the pre-fit values.

7.8.2 Statistical fluctuations of the $t\bar{t}+\geq 1b$ two-point systematics

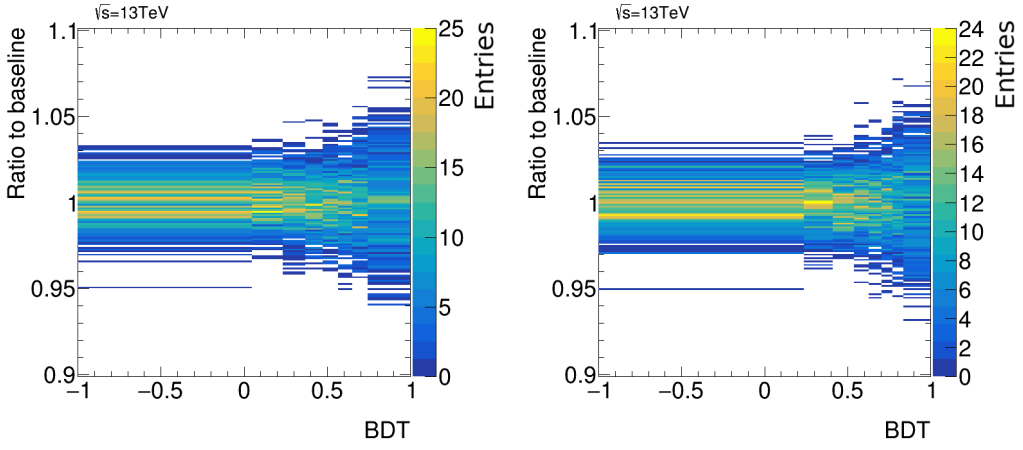


(a) (b)

Figure 7.14: Distributions of the ratio between the baseline $t\bar{t}+\geq 1b$ MadGraph5_aMC@NLO +Pythia8 sample and the bootstrap samples, displayed as a function of the Classification BDT for the two signal regions, (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ and (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$.

To investigate the effect of statistical fluctuations on the shapes of the $t\bar{t}+\geq 1b$ NLO matching and Parton shower & hadronization systematic uncertainties, the two MC samples used to derive them are smeared using the bootstrap method: MadGraph5_aMC@NLO +Pythia8 $t\bar{t}$+jets sample (shortened as MG+Py8) for the $t\bar{t}+\geq 1b$ NLO match and the Powheg +Herwig7 $t\bar{t}$+jets (or simply Pow+Her7) for the $t\bar{t}+\geq 1b$ PS&had. 500 alternative MC toy samples are produced. How the different datasets vary from the unsmeared baseline distribution can be seen in figures 7.14 and 7.15 for MG+Py8 and Pow+Her7, respectively. It can be noted that the fluctuations of the MG+Py8 sample are slightly larger than for the Pow+Her7.

For each of the 500 variations, a fit is performed and distributions of important variables are studied. In an Asimov fit the main question is how the fluctuations impact the uncertainty on $\mu_{t\bar{t}H}$. This can be see in figure 7.16 for the MG+Py8 sample, where the effect on $\mu_{t\bar{t}H}$ is small and the standard deviation is around 0.03. One can notice that the peak is slightly asymmetric though the effect is small enough to validate the use of a



(a) (b)

Figure 7.15: Distributions of ratio between the baseline $t\bar{t}+\geq 1b$ POWHEG +HERWIG7 sample and the bootstrap samples, displayed as a function of the Classification BDT for the two signal regions, (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ and (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$.

Gaussian distribution to estimate the uncertainties.

Another important feature in the figure is that the baseline, or the value from the fit of the nominal unsmeared model, is approximately in the center of the peak. During the optimization of this analysis, when the TTRES smoothing was used instead of the PARABOLIC for the modeling systematic uncertainties (see section 6.7.2), the smoothing introduced a bias in the shape and lead to the baseline being several standard deviations away from the center of the peak. This can be seen in figure 7.17, which shows the same distribution as figure 7.16 but with the older smoothing algorithm used on the $t\bar{t}+\geq 1b$ NLO match systematic. This further strengthens the confidence in the PARABOLIC smoothing.

The effect of the smearing on other variables can be found in table 7.2 for the MG+PY8 sample and in table 7.3 for POW+HER7. For both samples it shows that their impact on the constraint of the systematic they are used to derive (Up and down values in the table) is around 0.04. Since the $t\bar{t}+\geq 1b$ PS is not correlated to $\mu_{t\bar{t}H}$, the impact of POW+HER7 is negligible, while for th MG+PY8 it is around 3% as mentioned previously.

The impact on the pulls and constraints of the systematic variations and the effect on the normalization of $t\bar{t}+\geq 1b$ is further studied in the background-only fit to the blinded data. The results can be found in tables 7.2 and 7.3 for the MG+PY8 and the POW+HER7 sample, respectively. For the former, the impact on the $k(t\bar{t}+\geq 1b)$ is negligible. The $t\bar{t}+\geq 1b$ NLO match systematic itself has relatively large fluctuations and the central value shifts by around 0.11 and constraint by 0.04. Since the systematic is strongly correlated to the signal, the effect of MG+PY8 is later investigated in the full fit to the data.

The $t\bar{t}+\geq 1b$ Parton shower & hadronization systematic is more correlated to the $t\bar{t}+\geq 1b$ normalization, but effects of the smearing are still smaller than 0.01. Fluctuations of the central value of the systematic uncertainty are of order 0.11, where the constraint varies within 0.04, similar values to the $t\bar{t}+\geq 1b$ NLO match. This is not surprising as the two samples have comparable statistics.

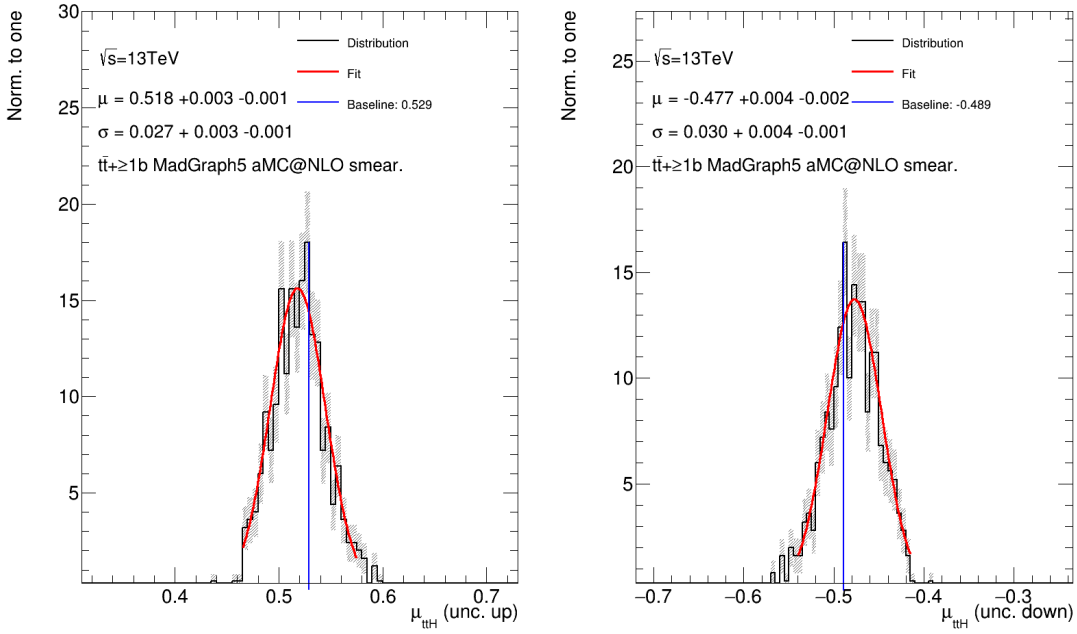


Figure 7.16: Distribution of the up (left) and down (right) uncertainty on $\mu_{t\bar{t}H}$ for different MC bootstrap samples of the $t\bar{t}+\geq 1b$ MadGraph5_aMC@NLO+Pythia8 sample. The distribution, representing the impact of the statistical fluctuations of the sample, is fitted by a Gaussian function (in red) to estimate the standard deviation σ of the statistical fluctuations. The baseline (in blue) represents the value given by the nominal unsmeared sample. The gray shading represents statistical uncertainties.

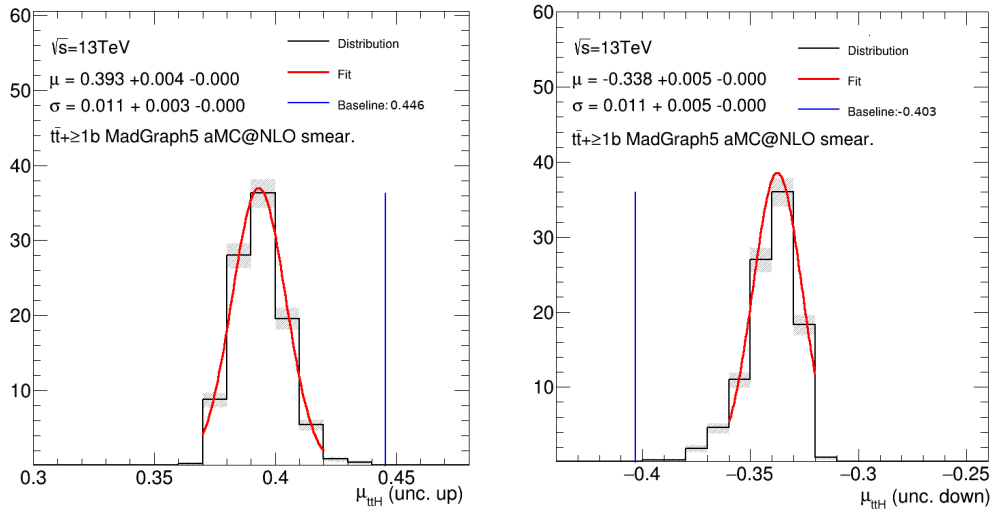


Figure 7.17: Distributions of the up (left) and down (right) uncertainty on μ for different MC bootstrap samples of the $t\bar{t}+\geq 1b$ MadGraph5_aMC@NLO+Pythia8 sample with an older smoothing. The distribution, representing the impact of the statistical fluctuations of the sample, is fitted by a Gaussian function (in red) to estimate the standard deviation σ of the statistical fluctuations. The baseline (in blue) represents the value given by the nominal unsmeared sample. The gray shading represents statistical uncertainties.

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
ASIMOV (full)	$\mu_{t\bar{t}H}$	-	0.027	0.030
	k($t\bar{t}+\geq 1b$)	-	< 0.01	< 0.01
	$t\bar{t}+\geq 1b$ NLO match.	-	0.044	0.044
BONLY (blinded)	k($t\bar{t}+\geq 1b$)	< 0.01	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.014	< 0.01	< 0.01
	$t\bar{t}+\geq 1b$ NLO match.	0.112	0.039	0.039

Table 7.2: Statistical impact of the MadGraph5_aMC@NLO+Pythia8 $t\bar{t}+\geq 1b$ sub-component on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the Asimov and background-only (BONLY) type fit. The up and down values refer to the absolute uncertainties for the free-floating normalization parameters $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ and for the other Nuisance parameters they refer to the constraints expressed in units of the pre-fit systematic uncertainty.

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
ASIMOV (full)	$\mu_{t\bar{t}H}$	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1b$)	-	< 0.01	< 0.01
	$t\bar{t}+\geq 1b$ PS	-	0.021	0.021
BONLY (blinded)	k($t\bar{t}+\geq 1b$)	< 0.01	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.024	0.015	0.08
	$t\bar{t}+\geq 1b$ PS	0.113	0.036	0.036

Table 7.3: Statistical impact of the Powheg+Herwig7 $t\bar{t}+\geq 1b$ sub-component on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the Asimov and background-only (BONLY) type fit. The up and down values refer to the absolute uncertainties for the free-floating normalization parameters $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ and for the other Nuisance parameters they refer to the constraints expressed in units of the pre-fit systematic uncertainty.

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
ASIMOV (full)	$\mu_{t\bar{t}H}$	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1b$)	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	-	< 0.01	< 0.01
	$t\bar{t}+\geq 1b$ NLO match	-	0.02	0.02
	$t\bar{t}+\geq 1b$ PS&had	-	0.02	0.02
	$t\bar{t}+\geq 1b$ ISR	-	0.02	0.02
	$t\bar{t}+\geq 1b$ FSR	-	0.09	0.09
	b-tag B0	-	< 0.01	< 0.01
BONLY (blinded)	k($t\bar{t}+\geq 1b$)	0.02	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.08	0.02	0.04
	$t\bar{t}+\geq 1b$ NLO match	0.10	0.04	0.04
	$t\bar{t}+\geq 1b$ PS&had	0.10	0.04	0.04
	$t\bar{t}+\geq 1b$ ISR	0.10	0.03	0.03
	$t\bar{t}+\geq 1b$ FSR	0.33	0.13	0.13
	b-tag B0	0.01	< 0.01	< 0.01

Table 7.4: Statistical impact of the nominal POWHEG+PYTHIA8 $t\bar{t}b\bar{b}$ sample on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the Asimov and background-only (BONLY) type fit. The up and down values refer to the absolute uncertainties for the free-floating normalization parameters $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ and for the other Nuisance parameters they refer to the constraints expressed in units of the pre-fit systematic uncertainty.

7.8.3 Statistical fluctuations of the nominal $t\bar{t}b\bar{b}$ sample

Since the $t\bar{t}+\geq 1b$ ISR and FSR systematics are derived from the nominal $t\bar{t}b\bar{b}$ sample through application of additional weights, the statistical fluctiations cannot be investigated separately from the nominal. Instead, the nominal POWHEGBOX+PYTHIA8 $t\bar{t}b\bar{b}$ sample and all its correlated systematics are smeared in a correlated manner, all of them using the same bootstrap weight per event. To evaluate the statistical fluctuations, one therefore has to produce bootstrap samples for all the correlated systematic variations, leading to larger computational requirements. For this reason only 100 alternative toy samples were produced for each.

The results are summarized in table 7.4. For the Asimov fit, the impact on $\mu_{t\bar{t}H}$ is again found to be less than 1%. The impact on the $t\bar{t}+\geq 1b$ normalization in the background-only fit is around 2%, where the distribution of the values for different bootstrap samples can be found in figure 7.18.

The central value of the $t\bar{t}+\geq 1b$ ISR systematic is biased by up to 10% and the constraint by 3%, a comparable performance to the $t\bar{t}+\geq 1b$ parton shower. The FSR has large fluctuations of the systematic weights and is therefore much more impacted by the statistics of the nominal sample. The systematic pull is fluctuating by around 33% and the constraint by 13%. Despite the worse performance of the FSR systematic, the overall impact on the measured $t\bar{t}H$ and $t\bar{t}+\geq 1b$ normalizations is small.

7.8.4 Statistical fluctuations of the signal samples

The same study of statistical effects was done for the $t\bar{t}H$ signal and its systematics. The impact was found to be minimal, for most of the distributions are smaller than 0.1%.

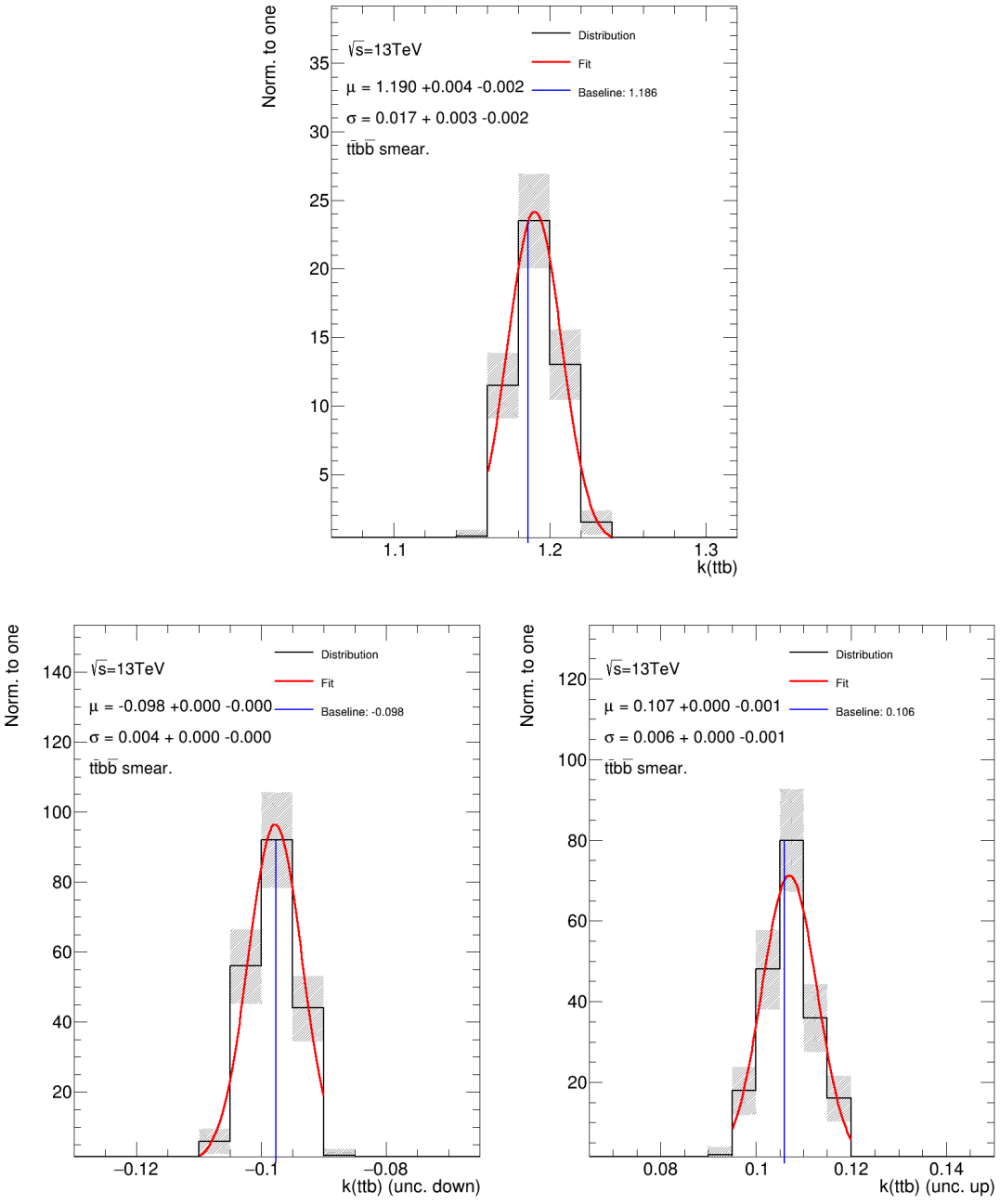


Figure 7.18: Distributions of the $t\bar{t}+\geq 1b$ normalization $k(t\bar{t}+\geq 1b)$ (top) and its uncertainty (bottom) for different MC toys of the nominal PowhegBox+Pythia8 $t\bar{t}b\bar{b}$ sample. The distribution is fitted by a Gaussian to estimate the standard deviation of the statistical fluctuations. The gray shading represents statistical uncertainties.

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
ASIMOV (full)	$\mu_{t\bar{t}H}$	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	-	0.01	< 0.01
	$t\bar{t}+\geq 1c$ PS&had	-	0.01	0.01
BONLY (blinded)	k($t\bar{t}+\geq 1b$)	< 0.01	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.14	0.10	0.10
	$t\bar{t}+\geq 1c$ PS&had	0.18	0.05	0.05

Table 7.5: Statistical impact of the Powheg+Herwig7 $t\bar{t}+\geq 1c$ sub-component on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the Asimov and background-only (BONLY) type fit. The up and down values refer to the absolute uncertainties for the free-floating normalization parameters $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ and for the other Nuisance parameters they refer to the constraints expressed in units of the pre-fit systematic uncertainty.

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
ASIMOV (full)	$\mu_{t\bar{t}H}$	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	-	< 0.01	< 0.01
	$t\bar{t}+\geq 1c$ NLO match	-	0.04	0.04
BONLY (blinded)	k($t\bar{t}+\geq 1b$)	< 0.01	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.03	0.02	0.01
	$t\bar{t}+\geq 1c$ NLO match	0.34	0.06	0.06

Table 7.6: Statistical impact of the MadGraph5_aMC@NLO+Pythia8 $t\bar{t}+\geq 1c$ sub-component on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the Asimov and background-only (BONLY) type fit. The up and down values refer to the absolute uncertainties for the free-floating normalization parameters $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ and for the other Nuisance parameters they refer to the constraints expressed in units of the pre-fit systematic uncertainty.

The biggest effect is on the uncertainty of the signal strength $\mu_{t\bar{t}H}$ when the nominal $t\bar{t}H$ sample is smeared, where it is 0.2%. Hence, it is safe to assume that finite statistics of the $t\bar{t}H$ samples will not have any impact on the result of the measurement.

7.8.5 Statistical fluctuations of other $t\bar{t}$+jets components

Statistical effects of the $t\bar{t}+\geq 1c$ component of the MadGraph5_aMC@NLO+Pythia8 and Powheg+Herwig7 samples can be found in tables 7.5 and 7.6, respectively. The former has a small impact in general except for the central value of the $t\bar{t}+\geq 1c$ NLO match systematic, which fluctuates by about 0.34. This partially shows that the pull on this systematic does not have a large impact on the fit result. The impact of the Pow+Her7 sample is the most importantly on the $t\bar{t}+\geq 1c$ normalization, where it is almost 0.15, with a 0.10 variation of the constraints.

Variations of the nominal sample can be found in table 7.7. The impact shown there is minimal, the largest fluctuations being the NLO match and FSR systematic. The $t\bar{t}$+light component was tested as well, but the impact is generally small. Only a 0.12 effect on the $t\bar{t}+\geq 1c$ normalization is noteworthy.

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
ASIMOV (full)	$\mu_{t\bar{t}H}$	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1b$)	-	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	-	< 0.01	< 0.01
	$t\bar{t}+\geq 1c$ NLO match	-	0.02	0.02
	$t\bar{t}+\geq 1c$ PS&had	-	0.01	0.01
	$t\bar{t}+\geq 1c$ ISR	-	< 0.01	< 0.01
	$t\bar{t}+\geq 1c$ FSR	-	0.05	0.05
	b-tag C0	-	< 0.01	< 0.01
BONLY (blinded)	k($t\bar{t}+\geq 1b$)	0.01	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.09	0.04	0.07
	$t\bar{t}+\geq 1c$ NLO match	0.25	0.04	0.04
	$t\bar{t}+\geq 1c$ PS&had	0.14	0.05	0.05
	$t\bar{t}+\geq 1c$ ISR	0.09	0.01	0.01
	$t\bar{t}+\geq 1c$ FSR	0.25	0.08	0.08
	b-tag C0	0.05	0.04	0.04

Table 7.7: Statistical impact of the nominal POWHEG+PYTHIA8 $t\bar{t}+\geq 1c$ sub-component on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the Asimov and background-only (BONLY) type fit. The up and down values refer to the absolute uncertainties for the free-floating normalization parameters $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ and for the other Nuisance parameters they refer to the constraints expressed in units of the pre-fit systematic uncertainty.

7.9 Results

Finally, results of the fit to the data in the full single lepton phase-space are presented here. The resulting value of the signal strength is:

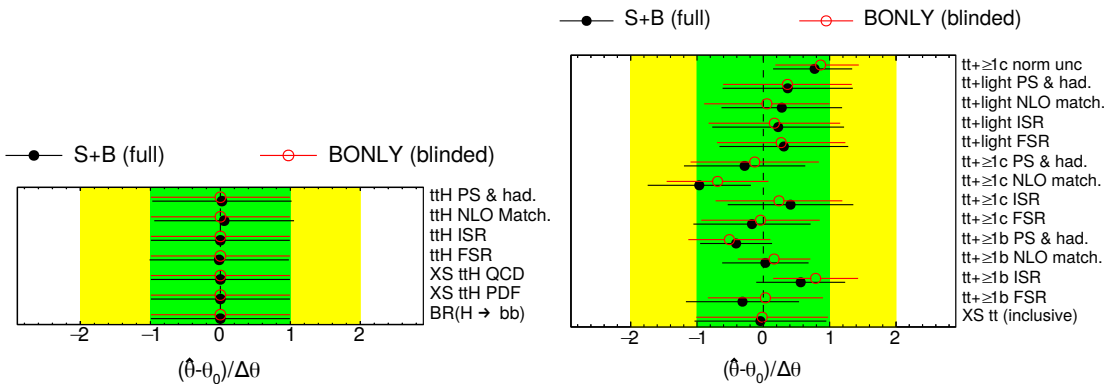
$$\mu_{t\bar{t}H}^{\text{single lepton}} = 0.54^{+0.61}_{-0.58} = 0.54^{+0.25}_{-0.25}(\text{stat.})^{+0.55}_{-0.52}(\text{syst.})$$

corresponding to a measured significance of 0.9σ. The significance is lower than the expected value of 1.8σ due to the low measured value of $\mu_{t\bar{t}H}$. The value of the $t\bar{t}+\geq 1b$ normalization is measured as $k(t\bar{t}+\geq 1b) = 1.17^{+0.10}_{-0.10}$, in agreement with the background-only value of 1.16 as described in section 7.5.

The value of $\mu_{t\bar{t}H}$, though low, is still in agreement with the Standard Model prediction. Whether this is only a statistical fluctuation or a consequence of background mis-modeling is studied more in the next chapter in the context of a combination with the dilepton channel, where the additional regions help to mitigate some of the shortcomings of the single-lepton channel.

The nuisance parameters of the $t\bar{t}$+jets and $t\bar{t}H$ systematic uncertainties and their constraints are displayed in figure 7.19 for the full fit, where they are directly contrasted to the background-only fit in the blinded data. Overall, the differences between the two fits are small. This seems to suggest that the accuracy of the background modeling does not differ significantly in the previously blinded bins. That is not surprising, since they contribute only a small part of the overall statistics. It is, however, possible that the remaining differences are absorbed by the signal rather than the modeling systematics.

The post-fit comparison of the data and the Monte Carlo prediction can be found in figure 7.20. It does not show any significant mis-modeling. Both 5 jet regions have a bin



(a) (b)

Figure 7.19: Resulting pulls and constraints of the $t\bar{t}H$ (a) and $t\bar{t}$+jets (b) uncertainties, comparing the background-only fit to the blinded data (BONLY) and signal+background fit to the whole phase-space (S+B).

Fit type	Parameter	Statistical impact		
		Central val.	Up value	Down value
S+B (full)	$\mu_{t\bar{t}H}$	0.121	0.039	0.043
	k($t\bar{t}+\geq 1b$)	< 0.01	< 0.01	< 0.01
	k($t\bar{t}+\geq 1c$)	0.013	< 0.01	< 0.01
	$t\bar{t}+\geq 1b$ NLO match.	0.198	0.052	0.052

Table 7.8: Statistical impact of the MG+Py8 $t\bar{t}+\geq 1b$ sub-component on various parameters of the fit model, specifically for their central value and the up and down uncertainties. Values are shown for the full fit to the data.

where the prediction still differs from the measured data, which, however, does not point at any trend and can be attributed to larger statistical fluctuations. The goodness of fit probability, evaluated using the saturated model[140], is 69%.

As was described in section 7.1.4, the significance is derived by comparing post-fit values of the likelihood between a fit with and without the signal, the latter corresponding to a background-only fit in the full range. The post-fit values of the $t\bar{t}$+jets nuisance parameters of the two fits are displayed in figure 7.21, showing that the disparity between the two fits is driven mainly by a pull on the $t\bar{t}+\geq 1b$ NLO match systematic variation. The difference between the other parameters is small.

Finally, the impact of the limited statistics of the MG+Py8 $t\bar{t}+\geq 1b$ sample , discussed previously in section 7.8, is revisited for the full fit. The results can be found in table 7.8, showing a 20% variation of the $t\bar{t}+\geq 1b$ NLO match systematic and 12% effect on the signal strength $\mu_{t\bar{t}H}$, a small effect compared to the overall 60% uncertainty. This represents a significant improvement compared to the previous analysis where the statistical fluctuations played a larger role (see section 6.1).

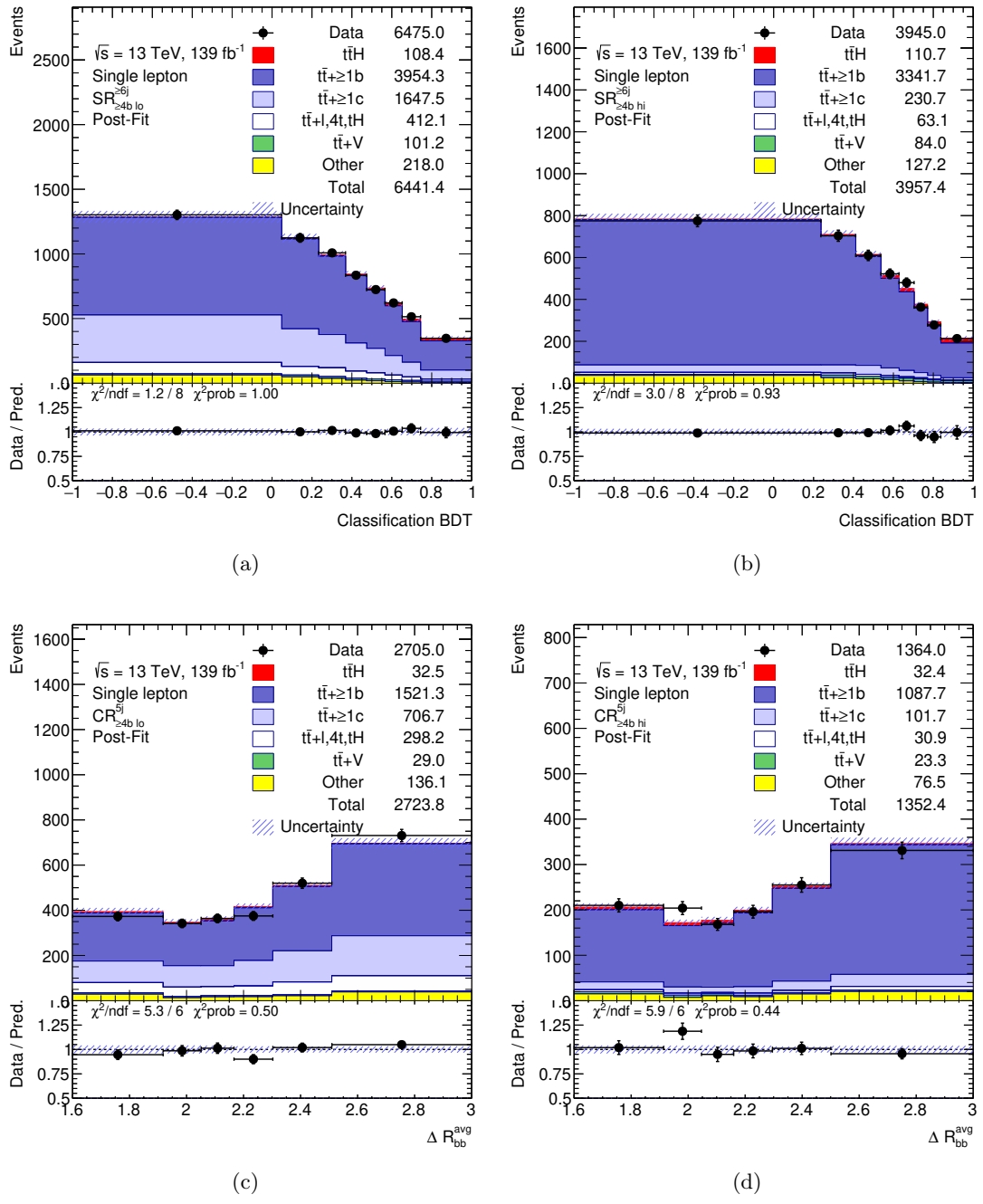


Figure 7.20: Post-fit modeling of the four analysis regions, (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$, (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$, (c) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{lo}}$ and (d) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{hi}}$, after performing the full signal+background fit, displayed as function of variable used, $\Delta R^{\mathrm{avg}}_{bb}$ for the 5 jet regions and the Classification BDT for 6 jet regions.

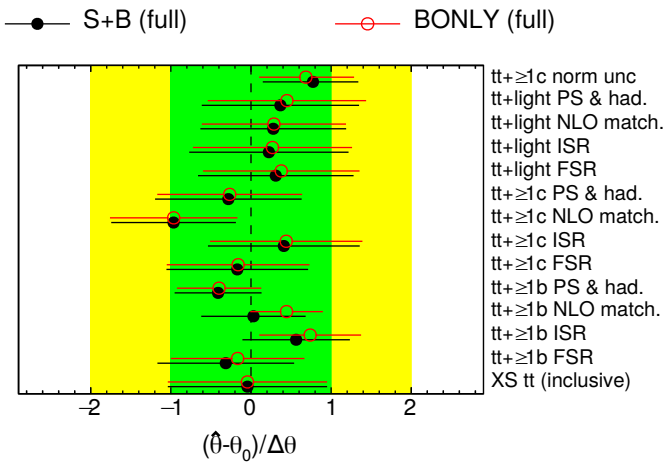


Figure 7.21: Resulting pulls and constraints of the $t\bar{t}$+jets uncertainties, comparing a background-only fit (BONLY) to a signal+background fit (S+B), both performed in the full analysis phase-space.

CHAPTER 8

Combination of the leptonic $t\bar{t}H(b\bar{b})$ channels

The results measured in the single lepton channel show a lower, though still statistically compatible, value of the signal strength with respect to the SM expectation. Both the background modeling and estimation of the signal can be improved by including additional $t\bar{t}H(b\bar{b})$ regions, rich in $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ respectively.

Additional $t\bar{t}H(b\bar{b})$ channels are discussed in the beginning of the chapter in section 8.1. Sections 8.1.2 to 8.4 are dedicated to the combination with the dilepton channel, where some additional studies performed on the combined model before the unblinding are discussed as well. Finally, results of the combined fit are introduced in section 8.5. Its interpretation and comparison to previous measurement is provided in section 8.6.

8.1 Additional analysis channels

Besides the single lepton channel, two other final states of the $t\bar{t}H(b\bar{b})$ production are possible, the dilepton and the all-hadronic containing two or no lepton, respectively. In terms of most of the main properties, the dilepton channel is the most similar to the single lepton channel. It has comparable background composition, making the combination quite straightforward.

The leptonic channels can be pre-selected by using leptonic triggers. In contrast, the all-hadronic channel has only jets in the final state and can be triggered only by the presence of jets, leading to a large multi-jet background. Such background is difficult to model with the generated samples and data-driven methods are used to estimate it instead. The resulting model has still significant uncertainties.

The single lepton channel is further divided between **resolved** channel, where the jets from the Higgs decay are well separated, and **boosted** channel, where the Higgs boson has a higher p_{T} and the decay products of different primary jets start to overlap and it becomes more practical to consider them as a single object: a large jet [146]. The boosted signature is not studied in the dilepton channel due to a much smaller expected yield.

The boosted channel uses looser requirements on the b-tagging to increase the statistics, leading to a higher contribution of the $t\bar{t}+\geq 1c$ and $t\bar{t}+$light components. Furthermore, it has a higher contribution of non-$t\bar{t}+$jets background, especially from $W+t$ and $W+$jet production. Finally, the boosted frame leads to different kinematics compared to the resolved channels.

Neither the boosted nor the all-hadronic channel are included in the combination since

Variable	Single lepton	Dilepton
Leading lepton p_{T}	> 27 GeV	
$N_{\text{lepton}}^{p_{\mathrm{T}}>10\text{ GeV}}$	$= 1$	$= 2$
N_{jet}	$>= 5$	$>= 3$
$N_{\text{b-jet}}^{70\%\text{ WP}}$	$>= 4$	
$N_{\text{hadr. tau}}$	$= 0$	$<= 1$
Other		$m_{ee/\mu\mu} > 15$ GeV Veto $83 < m_{ee/\mu\mu} < 99$ GeV

Table 8.1: Preselection of the single lepton and dilepton channel, based on objects defined in section 6.5.1: number of leptons with $p_{\mathrm{T}} > 10$ GeV $N_{\text{lepton}}^{p_{\mathrm{T}}>10\text{GeV}}$, number of jets N_{jet}, number of b-jets $N_{\text{b-jet}}^{70\%\text{ WP}}$ and number of hadronic tau leptons $N_{\text{hadr. tau}}$.

they were still in processes of development at the time of writing of this thesis. The dilepton channel is being summarized in an another thesis[147] and the combination presented in this chapter is a result of a collaboration with its author. The single lepton channel will from now on refer exclusively to the resolved channel, which was presented in the previous chapters.

8.1.1 Dilepton channel

The dilepton channel has some small differences in the nominal model compared to the single lepton channel, mainly some additional uncertainties on the $t\bar{t}Z$ background. Furthermore, the dilepton channel has a larger contribution of a background from misidentified lepton, because of a lower p_{T} threshold of the second lepton. However, neither has a significant impact on the measurement. The dilepton channel uses the same definition of reconstructed objects as the single lepton channel, previously described in section 6.5.1.

As for the constitution of the final state, the only difference with respect to the single lepton channel is that both W bosons from the top quark decay leptonically. Hence, there are two leptons in the final state but only 4 jets, all containing B hadrons. The event selection requires the two leptons to have an opposite charge, one with $p_{\mathrm{T}} > 27$ GeV and the second lepton with $p_{\mathrm{T}} > 10$ GeV. In the case where both leptons have the same flavor, there is a veto on events with invariant dilepton mass under 15 GeV and within 83-99 GeV to suppress background from low-mass (e.g. J/Ψ or Υ) and Z resonances. To avoid overlap with other $t\bar{t}H$ analyses, events with at least one hadronic tau lepton candidate are removed. The preselection is summarized and compared to the single lepton channel in table 8.1.

Analysis regions

The dilepton analysis has two signal regions modeled similarly to the single lepton channel (see section 6.5.2). They require at least four jets in the final state, the first region having a tight b-tagging requirement $\geq 4b$@60% and the second having a loose $\geq 4b$@70% selection excluding the tighter events. They are designated $\text{SR}_{\geq 4b\text{ hi}}^{\geq 4j}$ and $\text{SR}_{\geq 4b\text{ lo}}^{\geq 4j}$ respectively.

Two additional regions are defined for events with four or more jets, one with exactly 3 b-jets at the 60% working point ($\text{CR}_{3b\text{ hi}}^{\geq 4j}$) and one with exactly 3 b-jets at the 70% ($\text{CR}_{3b\text{ lo}}^{\geq 4j}$) with events from the tighter region excluded.

Finally, one more region is defined in events with exactly 3 jets in the final states, all of them tagged as b-jets at the 60% working point ($\text{CR}_{3b\text{ hi}}^{3j}$). The regions are summarized

in table 8.2.

Region	n_{lepton}	n_{jet}	$n_{\text{b-jet}}$ @60%	$n_{\text{b-jet}}$ @70%
$\text{SR}^{\geq 4j}_{\geq 4b\ \text{hi}}$	$= 2$	≥ 4	≥ 4	≥ 4
$\text{SR}^{\geq 4j}_{\geq 4b\ \text{lo}}$			< 4	
$\text{CR}^{\geq 4j}_{3b\ \text{hi}}$			≥ 3	≥ 3
$\text{CR}^{\geq 4j}_{3b\ \text{lo}}$			< 3	
$\text{CR}^{3j}_{3b\ \text{hi}}$		$= 3$	≥ 3	

Table 8.2: The definitions of the dilepton analysis regions, based on the number of jets n_{jet}, and the number of b-tagged jets $n_{\text{b-jet}}$ using the 60% and 70% working points. SR refers to signal regions and CR to control regions.

The background composition of the dilepton channels can be found in figure 8.1(a). As was already mentioned, the composition is not very different from that for the single lepton channel. Only the $\text{CR}^{\geq 4j}_{3b\ \text{lo}}$ has a larger contribution of the $t\bar{t}+\geq 1c$, which provides a better handle of this background.

The signal contribution and the statistical significance can be found in figure 8.1(b), which shows a similar fraction of the $t\bar{t}H$ in the tightest region (around 7%) but the statistical significance $S/\sqrt{B}$ is much lower due to an overall lower number of events.

Variables used in the fit

The dilepton analysis also uses a Boosted Decission Tree to increase the signal to background separation. A short description can be found in the MVA dedicated appendix A. This discriminant is also called a Classification BDT and is used in the signal regions.

In contrast to the single lepton channel, the control regions do not use any differential distribution and only the yield is fitted instead. The goal is to only obtain a better control over the fractions of the $t\bar{t}$+jets sub-components.

The pre-fit modeling of the signal regions can be found in figure 8.2, showing that the disagreement between the data and the Monte Carlo is mainly in normalization.

Results in the dilepton channel

In the Asimov fit, the dilepton channel measures $\mu_{t\bar{t}H} = 1.00^{+0.59}_{-0.54}$ which has a 5% larger uncertainty with respect to the single lepton channel, showing a slightly smaller sensitivity to the signal. In the fit to the data, the dilepton channel measures the following signal strength:

$$\mu^{\text{dilepton}}_{t\bar{t}H} = 1.43^{+0.69}_{-0.62},$$

a value larger than the Standard Model expectation $\mu_{t\bar{t}H} = 1$ but still compatible within the uncertainties. The dilepton channel also measures the value of the $t\bar{t}+\geq 1b$ normalization:

$$k(t\bar{t}+\geq 1b) = 1.22^{+0.09}_{-0.08},$$

which is comparable to the single lepton result, but with a smaller uncertainty. Comparison of the $k(t\bar{t}+\geq 1b)$ and of the $t\bar{t}$+jets nuisance parameters between the single lepton and dilepton channel can be found in figure 8.3. It shows large differences, especially for the $t\bar{t}+\geq 1b$ modeling. These are discussed in more detail in context of the background-only fit in section 8.3.

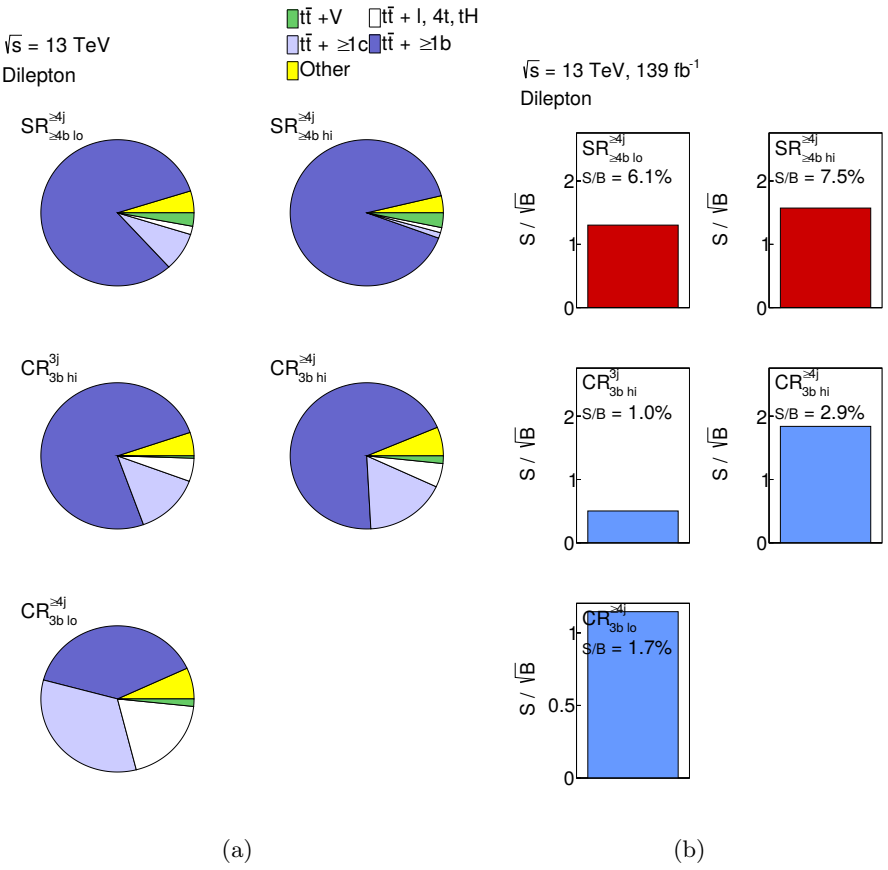
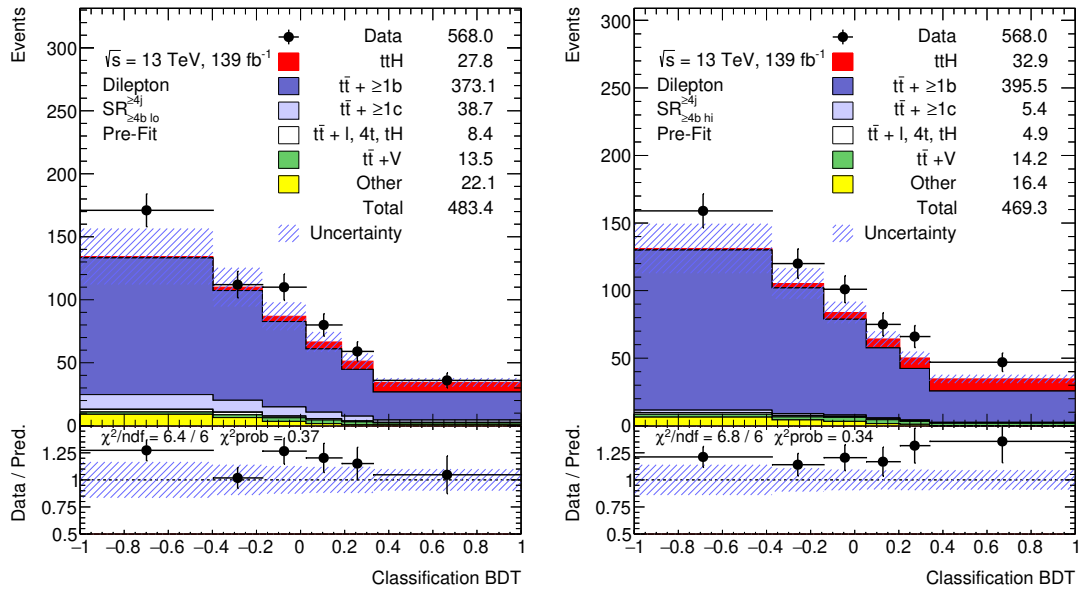
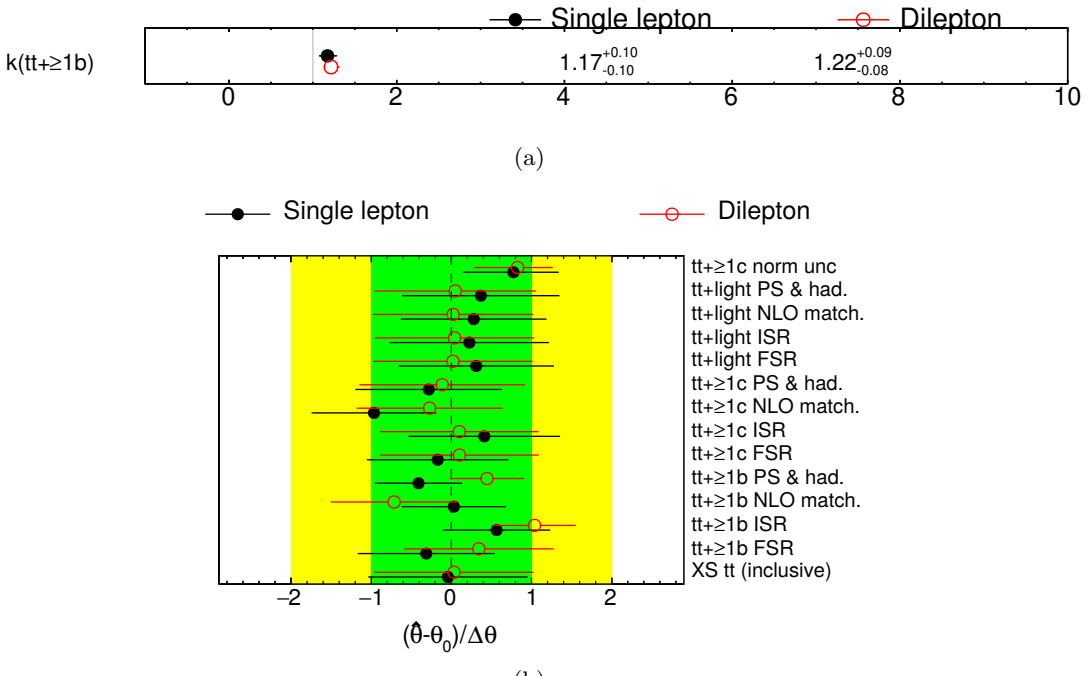


Figure 8.1: Background composition (a), and the signal over background ratio and the statistical significance (b) of the four analysis regions in the dilepton channel[147].



(a) (b)

Figure 8.2: Pre-fit modeling of the dilepton signal regions (a) $\mathrm{SR}^{\geq 4j}_{\geq 4b\ \mathrm{lo}}$ and (b) $\mathrm{SR}^{\geq 4j}_{\geq 4b\ \mathrm{hi}}$, displayed as a function of the discriminant of the classification BDT[147].



(a)

(b)

Figure 8.3: Resulting value and uncertainty on the $t\bar{t}+\geq 1b$ normalization (a) and pulls and constraints of the $t\bar{t}$+jets uncertainties (b) for a fit to the data comparing the single lepton and dilepton channel.

8.1.2 Combined analysis model

The two resolved leptonic channels, the single lepton and the dilepton, are combined in a single measurement. This increases the number of analyzed regions, leading to a better control over the various backgrounds. The definition of the analysis region of both channels is summarized in figure 8.4. All common systematic uncertainties are treated as correlated across the two channels. The nuisance parameters were previously summarized in section 7.2.

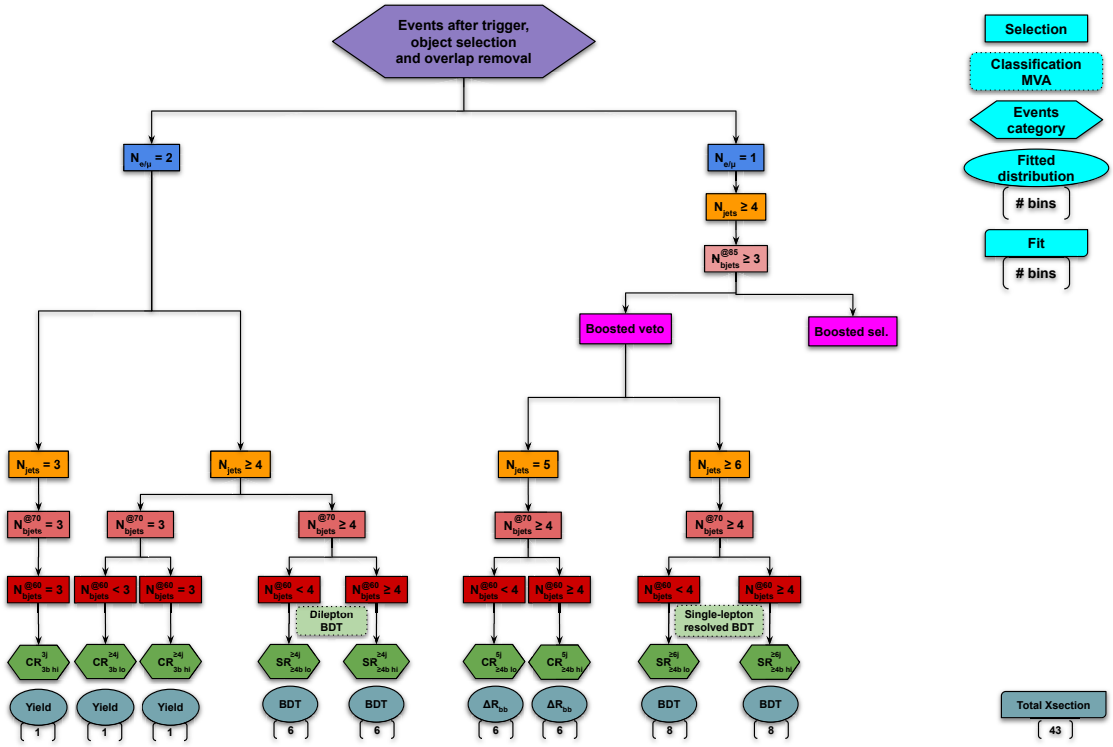


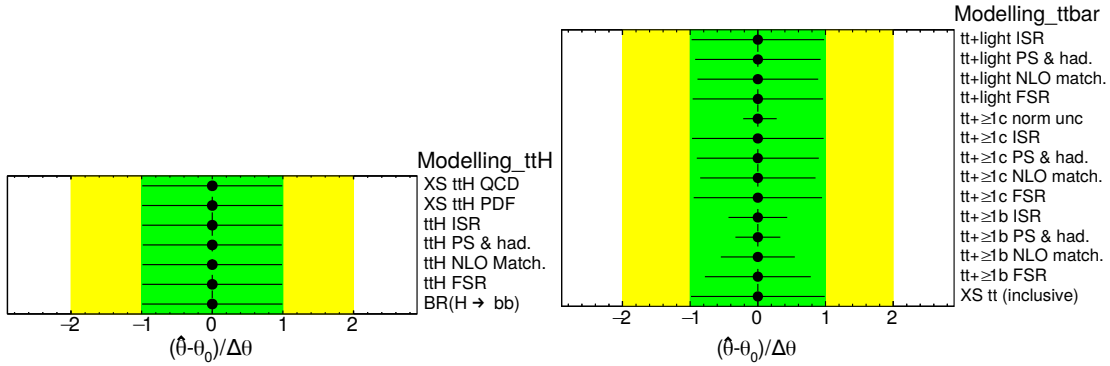
Figure 8.4: The analysis flow, showing the definition of channels, their regions and variables used in the fit. The number of bins in each region is shown as well. Courtesy of Timothée Theveneaux-Pelzer.

8.2 Asimov fit

The fit to the Asimov pseudodata was performed in the combination of the two channels, resulting in a signal strength $\mu_{t\bar{t}H} = 1.00^{+0.44}_{-0.40}$ with a statistical uncertainty of 20%. This corresponds to a significance of 2.6σ compared to the 2σ coming only from the single lepton channel. The uncertainty on the $t\bar{t}+\geq 1b$ normalization is around 7% with 1% statistical uncertainty.

The $t\bar{t}H$ and $t\bar{t}$+jets nuisance parameters are shown in figure 8.5. The $t\bar{t}+\geq 1c$ normalization uncertainty is reduced, as are the $t\bar{t}+\geq 1b$ modeling systematics. The $t\bar{t}+\geq 1c$ and $t\bar{t}$+light modeling uncertainties remain unconstrained.

The correlation matrix shows features similar to the single lepton channel, with the $t\bar{t}+\geq 1b$ NLO match being strongly correlated to the signal strength $\mu_{t\bar{t}H}$. There are small changes in correlation of the $t\bar{t}+\geq 1b$ ISR systematic and $k(t\bar{t}+\geq 1b)$ to $\mu_{t\bar{t}H}$ and $t\bar{t}+\geq 1b$ PS still remains highly correlated to the $t\bar{t}+\geq 1b$ NLO match.



(a) (b)

Figure 8.5: Resulting pulls and constraints of the $t\bar{t}H$ (a) and $t\bar{t}$+jets (b) uncertainties for fit to the Asimov dataset.

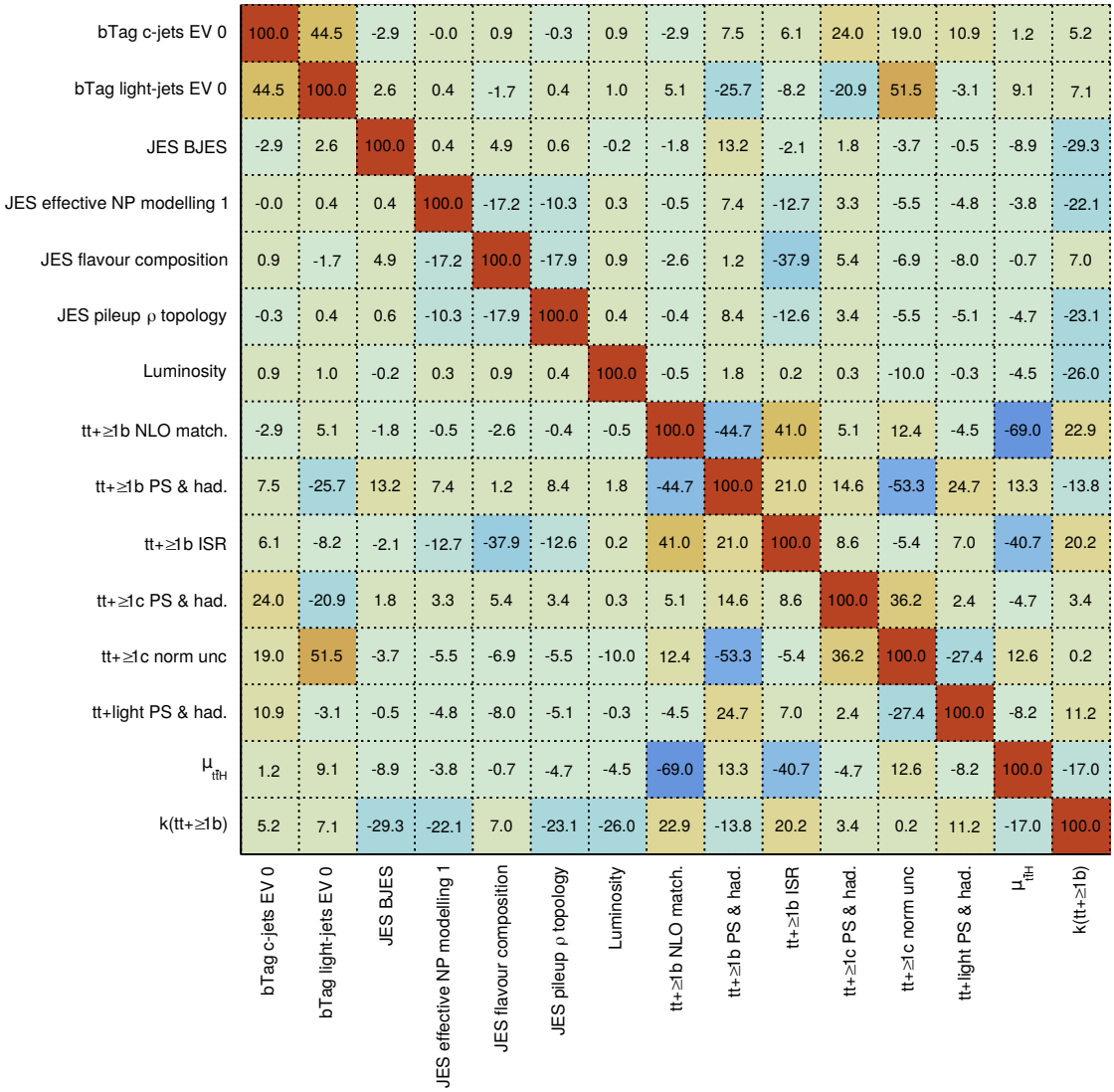


Figure 8.6: Correlation matrix for the fit to the Asimov dataset, showing all parameters which have at least one correlation larger than 20%.

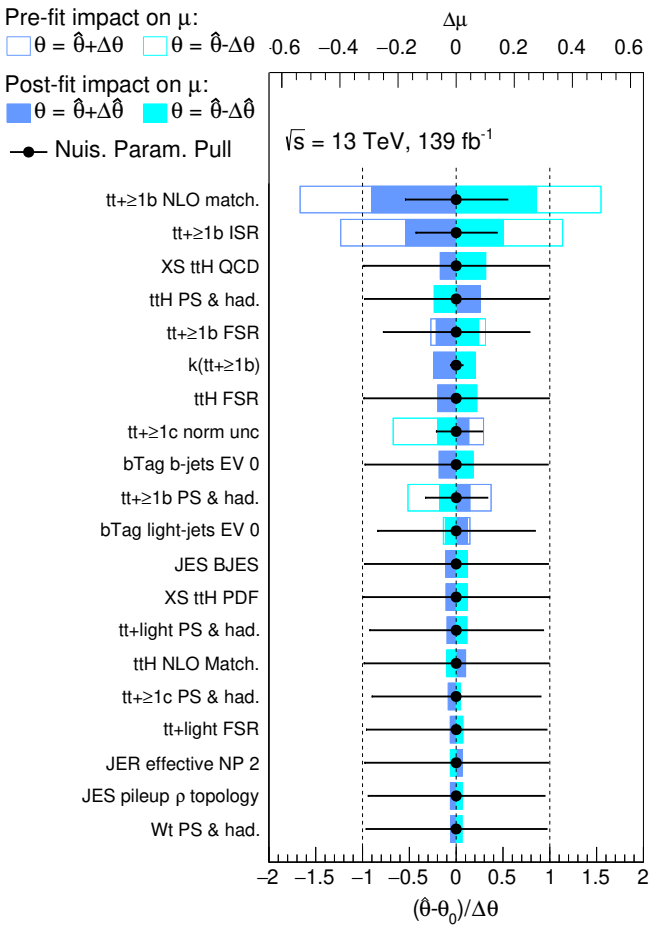


Figure 8.7: Ranking plot for 20 nuisance parameters with the largest impact on the parameter of interest $\mu_{t\bar{t}H}$ in combination of both channels. It is shown for the pre-fit uncertainty by the empty box and post-fit by the filled box. The post-fit shifts and constraints of the systematics are displayed as the black markers and the horizontal line respectively. Vertical dashed line correspond to ± 1 values of the nuisance parameters.

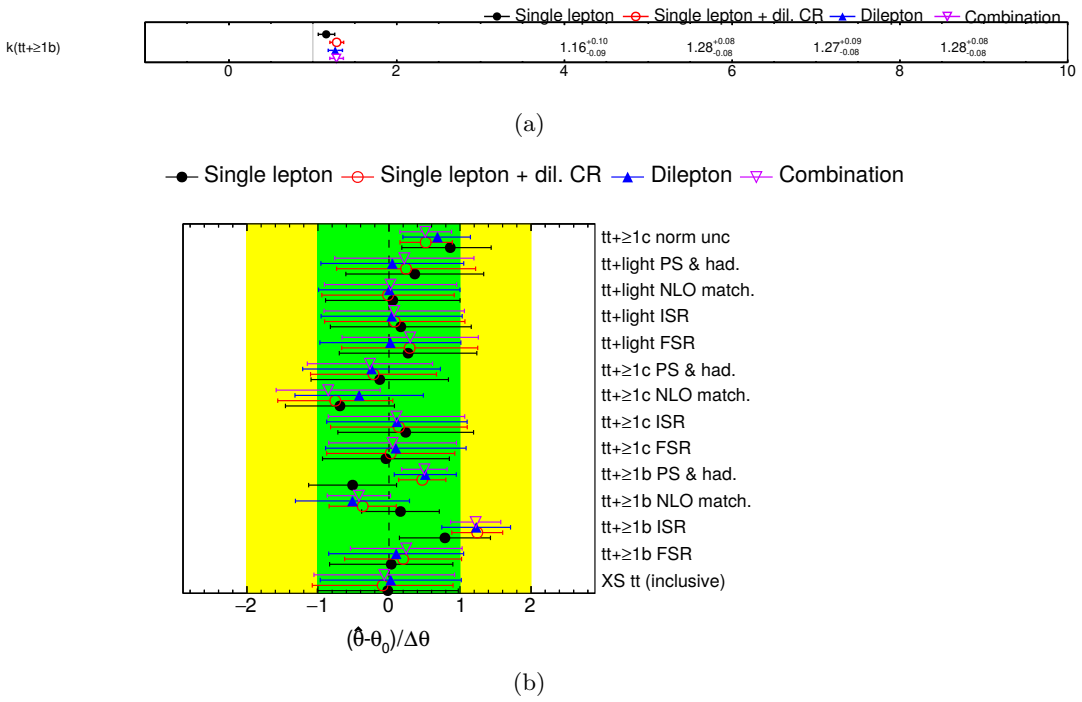


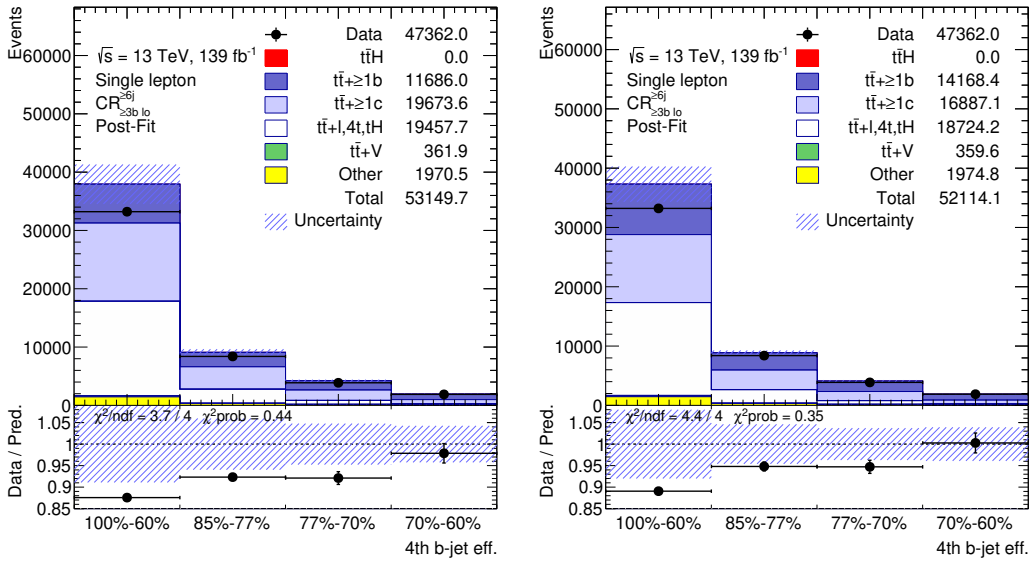
Figure 8.8: Resulting value and uncertainty on the $t\bar{t}+\geq 1b$ normalization (a) and pulls and constraints of the $t\bar{t}$+jets uncertainties (b) for a fit to the data in the revealed bins, comparing the single lepton and dilepton channel and the combination. In addition, the result of a fit in the single lepton channel with the dilepton control regions is shown.

The ranking plot in figure 8.7 reflects the correlation matrix, with the $t\bar{t}+\geq 1b$ NLO match and ISR systematic uncertainties, which are highly correlated to the $\mu_{t\bar{t}H}$, having significantly larger impact than any other systematic. Their understanding will be the most important for the resulting significance of the measurement.

8.3 Background-only fits to blinded data

The $t\bar{t}+\geq 1b$ normalization in the background-only fit, shown in figure 8.8(a), has around one sigma deviation in the combination with respect to the single lepton result. The $t\bar{t}+\geq 1b$ modeling uncertainties, displayed in figure 8.8(b), seem to be driven by the dilepton channel, in which they differ only slightly from the combination. On the other hand, the $t\bar{t}+\geq 1c$ modeling, with the exception of the $t\bar{t}+\geq 1c$ normalization discussed later on, is mainly driven by the single lepton channel. This is mainly because the dilepton control regions, which contain higher contribution of the $t\bar{t}+\geq 1c$ component, are not fitted differentially and thus only the normalization is measured precisely.

The difference in the resulting nuisance parameters between the two channels is probably caused by a badly measured $t\bar{t}+\geq 1c$ normalization in the single lepton channel and its connection to the $k(t\bar{t}+\geq 1b)$ and $t\bar{t}+\geq 1b$ ISR systematic. This is discussed more in the next section.The figure 8.8 also shows the single lepton channel combined with dilepton control regions. This is done to increase the contribution of the $t\bar{t}+\geq 1c$ component, which helps to constrain its normalization. The fit in such combination shows an excellent agreement with the full combination. This shows that the dilepton signal regions have a small effect on the measurement of $t\bar{t}+\geq 1b$ and $t\bar{t}+\geq 1c$ modeling.



(a) (b)

Figure 8.9: Distributions of the b-tagging category of the 4th b-jet (as in jet with 4th highest b-tagging discriminant) in an event selected in the single lepton channel with six jets and 3 b-jets at the 70% working point. The distributions are shown for a fit (a) in single lepton channel only and (b) for the combined result.

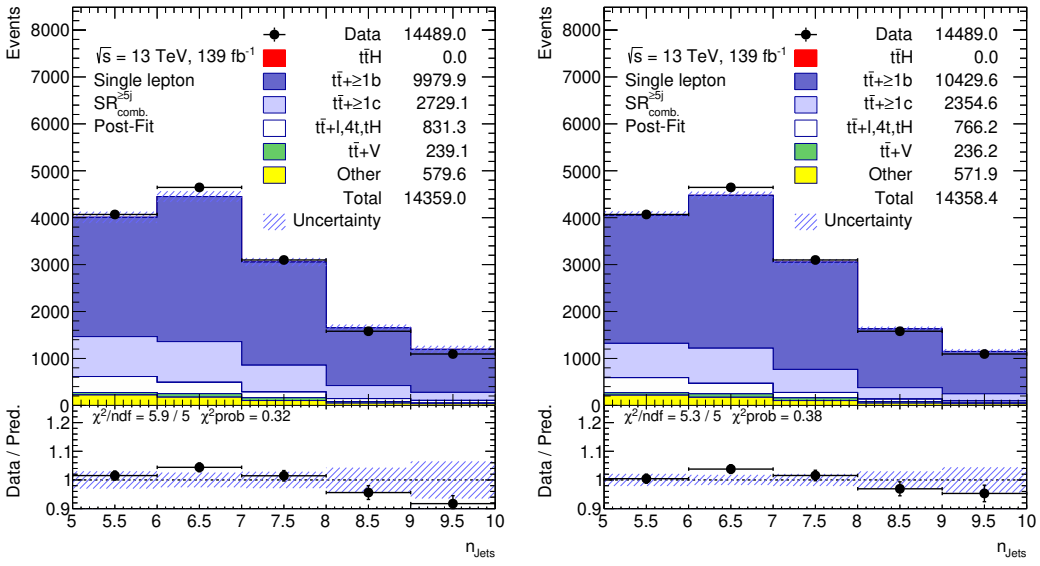
8.3.1 Validation of large parameter pulls

Two systematic uncertainties are shifted significantly from their nominal value in the combined fit while being constrained and thus should be investigated in more detail.

$t\bar{t}+\geq 1c$ normalization

An advantage of the inclusion of the dilepton channel is a region with a relatively large contribution of the $t\bar{t}+\geq 1c$ background process, which results in a better estimation of the $t\bar{t}+\geq 1c$ normalization. However, it is possible that the $t\bar{t}+\geq 1c$ process behaves differently in the single lepton channel and it is important to confirm that the fraction of $t\bar{t}+\geq 1c$ events in the single lepton channel is well-defined.

For this reason, $t\bar{t}+\geq 1c$ validation regions were investigated in a 3 b-jet phase-space. However, it contains a large fraction of $t\bar{t}$+light events, so a more complex selection is needed. Figure 8.9 shows distributions of the b-tagging working point of the fourth b-jet. The second and the third bin of the distribution, corresponding to the 85% and 77% working points of the b-tagging, have a larger fraction of the $t\bar{t}+\geq 1c$ component while containing almost no contribution from the $t\bar{t}$+light process. The figure compares the fit in the single lepton and for the combination of the two channels and though the difference is not large, it still suggests that modeling of the $t\bar{t}+\geq 1c$ fraction is better in the combined fit. The resulting value of the $t\bar{t}+\geq 1c$ normalization parameter is 1.5, close to the value 1.6 measured in the $t\bar{t}b\bar{b}$ analysis[122]. This is an improvement with respect to the single lepton channel which measured a value of 1.85.



(a) (b)

Figure 8.10: Distributions of the number of jets in a phase-space combining all single lepton regions. The distributions are shown for a fit (a) in single lepton channel only and (b) for the combined result.

$t\bar{t}+\geq 1b$ Initial State Radiation pull

A larger change in the combined fit is seen for the pull on the $t\bar{t}+\geq 1b$ ISR nuisance parameter. This systematic has a large impact on the jet multiplicity modeling. In the single lepton channel only two bins in the multiplicity are considered (the 5 and 6 jet regions). The fit then correct the multiplicity through the $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ normalization parameters, since the penalty in the likelihood is much small than a large pull on the $t\bar{t}+\geq 1b$ ISR systematic.

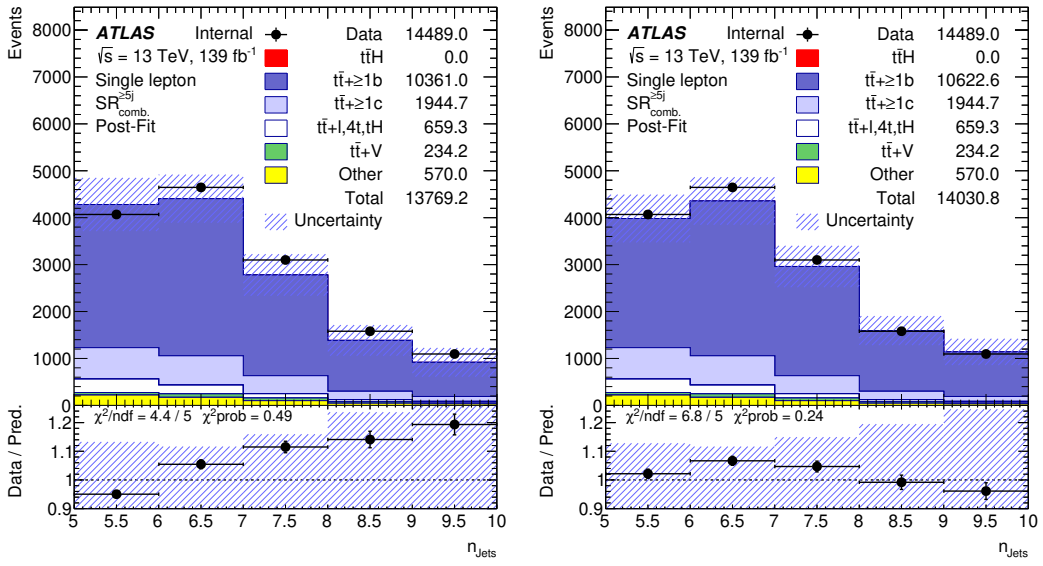
In the combined fit, the three and four jet dilepton regions are included in the phase-space, so modeling of the jet multiplicity has a larger impact. The improvement in the combination can be seen in figure 8.10 showing the jet multiplicity in the $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ region.

To demonstrate the effect more directly, figure 8.11 shows the same distribution, but comparing a situation where only post-fit values of the $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ normalizations are applied, compared to a case where the measured $t\bar{t}+\geq 1b$ ISR shift is applied as well. By contrasting these two plots, and comparing them back to the figure 8.10, it can be concluded that not only the $t\bar{t}+\geq 1b$ ISR systematic improves the agreement significantly.

This result points to a potential improvement in the nominal modeling of the $t\bar{t}b\bar{b}$ sample. The ISR systematic is dominated by a variation of the renormalization and factorization scale. Specifically, the $+1\sigma$ variation, which agrees with the data better than the central value in the observation, corresponds to a variation at a half of the nominal renormalization and factorization scales. Producing a sample with the central value at this lower scale should improve the agreement of the nominal sample with the data.

8.3.2 Data-driven expectation

A fit to the data-driven pseudodata, constructed as described in section 7.6, was performed in the combination to provide a more accurate prediction compared to the Asimov model.



(a) (b)

Figure 8.11: Distributions of the number of jets in a phase-space combining all single lepton regions. The distributions are shown (a) with application of only the $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ normalization factors and (b) with a $t\bar{t}+\geq 1b$ ISR systematic shift applied on top.

It results at a measured value $\mu_{t\bar{t}H} = 1.02^{+0.50}_{-0.45}$ with a significance 2.3σ, a value slightly lower than the 2.6σ predicted by the Asimov fit due to higher normalization of the backgrounds.

8.3.3 Impact of different smoothing methods

One of the tools of the analysis is the smoothing used in preparation of the histograms, described in detail in section 6.7. Since different smoothing methods lead to slightly different shapes in the histograms, it is important to check that choice of the algorithm in the $t\bar{t}H(b\bar{b})$ analysis does not significantly impact the results.

Two different models are compared to get an indication of a bias due to a choice of smoothing algorithm:

- Nominal: MaxVar is used for all systematics with the exception of the $t\bar{t}$+jets and the $t\bar{t}H$ modeling uncertainties, where Parabolic is used
- Alternative: MaxVar is used for all systematics with exception of the $t\bar{t}$+jets and the $t\bar{t}H$ modeling uncertainties. For the $t\bar{t}+\geq 1b$ NLO match, the $t\bar{t}H$ NLO match and the $t\bar{t}H$ PS&had systematic uncertainties the Parabolic smoothing is used. For the remainder of the systematics the TTRes method is used.

The motivation for the two schemes was given previously in section 6.7.

The background-only fit in the revealed bins is performed for the two smoothing schemes. The normalization factor of the $t\bar{t}+\geq 1b$, displayed in figure 8.12(a), shows a negligible difference between the two schemes.

A comparison of the pulls and the constraints of the $t\bar{t}$+jets modeling between the two schemes can be found in figure 8.12(b). It shows that differences between the two schemes are always much smaller than uncertainty of a given parameter. Hence, the choice of the smoothing algorithm seems to have a small impact on the measurement.

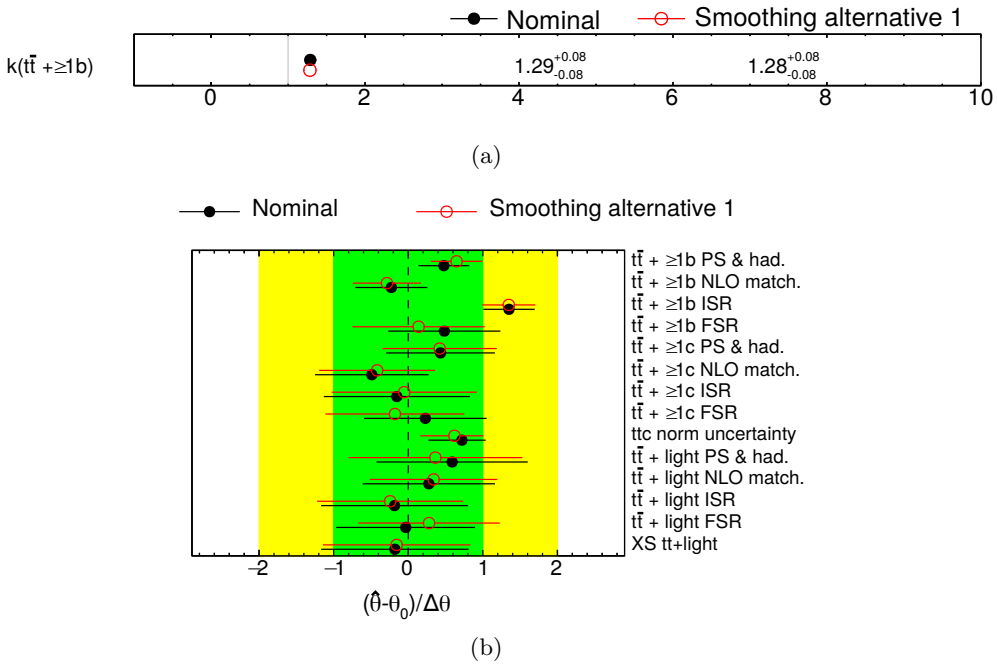


Figure 8.12: Resulting value and uncertainty on the $t\bar{t}+\geq 1b$ normalization (a) and pulls and constraints of the $t\bar{t}$+jets uncertainties (b) for a fit to the data in the revealed bins, comparing two smoothing stategies.

8.4 Pseudodata based on Sherpa generator

Similarly to the single lepton channel in section 7.7, the nominal model was tested in the combination on pseudodata created by replacement of the nominal Powheg +Pythia8 $t\bar{t}b\bar{b}$ sample with the Sherpa alternative. The results, summarized in figure 8.13, show similar features as in the single lepton only measurement.

The value of $\mu_{t\bar{t}H}$ measured in the data-driven pseudodata is biased by around 15%, while in the fit in the full range only by 5%. This discrepancy is driven by a different shift in the $t\bar{t}+\geq 1b$ NLO match systematic, which is pulled by around -0.5σ and strongly correlated to the signal. There is also a slight difference in the $k(t\bar{t}+\geq 1b)$, which is caused by its correlation to the NLO match systematic. In all cases the measured $\mu_{t\bar{t}H}$ values are compatible with the SM prediction.

8.5 Results of the measurement in the combined channel

A full fit of the nominal model to the data was performed for the combination of the two leptonic channels, resulting in the following values of the two free floating parameters:

$$\mu_{t\bar{t}H} = 0.84^{+0.45}_{-0.39}(\text{syst.}) \pm 0.21(\text{stat.})$$

$$k(t\bar{t}+\geq 1b) = 1.27 \pm 0.08,$$

with a measured significance of 1.9σ, smaller than the expected value of 2.3σ due to the lower value of $\mu_{t\bar{t}H}$. The measured value is still close to the SM expectation of $\mu_{t\bar{t}H}$ =1 than the fit in the single lepton channel only.

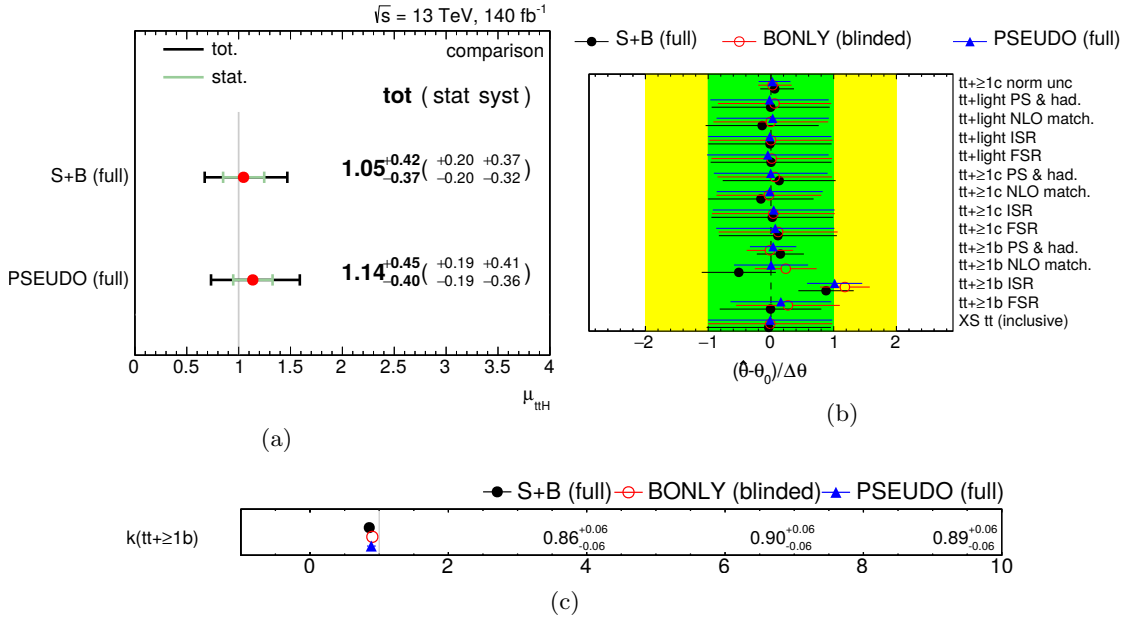


Figure 8.13: Summarizing plots of the three fits to Sherpa pseudodata: nominal fit, background-only fit in revealed bins and fit to data-driven pseudodata based on the results of the background-only fit. (a) shows values of the parameter of interest, (c) of the $t\bar{t}+\geq 1b$ normalization and finally (b) pulls and constraints of the $t\bar{t}$+jets modeling.

The measured $k(t\bar{t}+\geq 1b)$ is in an agreement with the expectation from the background-only fit, which gives $k(t\bar{t}+\geq 1b) = 1.28$. The post-fit values of the nuisance parameters, compared to the background-only value, can be found in figure 8.14.

The $t\bar{t}$+jets modeling does not change significantly for the $t\bar{t}+\geq 1c$ and $t\bar{t}$+light backgrounds between the two fits. In the $t\bar{t}+\geq 1b$ category, two differences can be observed. The ISR post-fit value is almost exactly at its $+1\sigma$ variation in the full fit while being slightly larger in the background-only fit. Similarly to the behavior observed in the pseudodata fits, the $t\bar{t}+\geq 1b$ NLO match systematic is also pulled towards larger negative values in the nominal fit.

The nuisance parameters of the experimental systematics show an excellent agreement between the two fits and no significant pull on the modeling of the $t\bar{t}H$ and non-$t\bar{t}+\geq 1b$ backgrounds is observed.

Post-fit modeling

The post-fit distributions of the single lepton regions can be found in figure 8.15, showing a good agreement between the data and the Monte Carlo prediction, especially in the 6 jet regions. More distributions for the single lepton channel can be found in appendix D.

Figure 8.16 shows the two signal regions of the dilepton channel. The post-fit agreement is not as good as in the single lepton channel: while the tighter region shows a deficit in the most signal like bins (which is why the measured value of $\mu_{t\bar{t}H}$ is higher in this channel), the looser region has a slightly opposite tendency, more in line with the single lepton channel. Which can be consequence of statistical fluctuations of the data in the last two bins of the tighter region. Nevertheless, the data and the MC distribution still agree within the uncertainties.

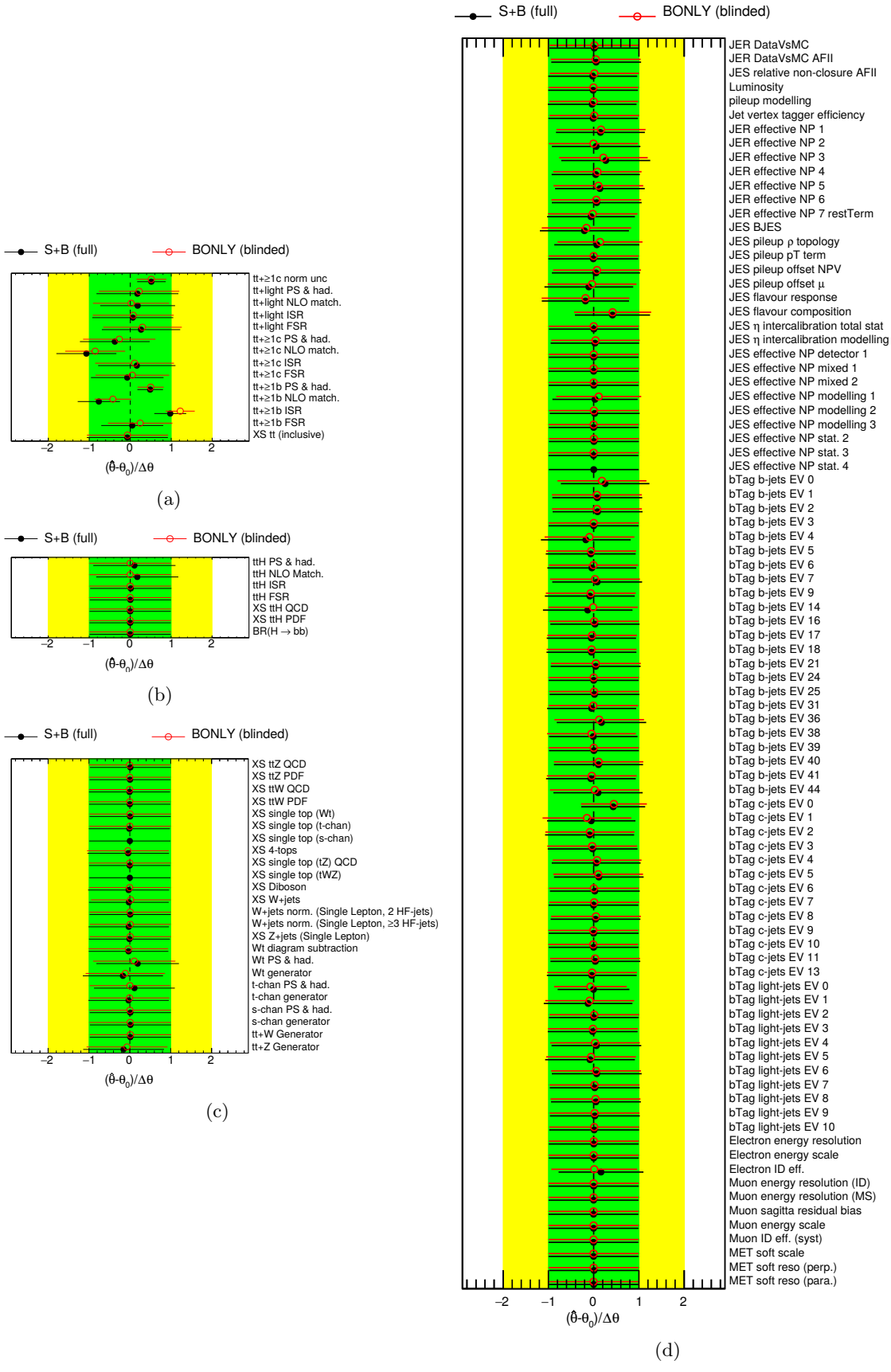


Figure 8.14: Resulting pulls and constraints of the $t\bar{t}$+jets modeling (a), the modeling the $t\bar{t}H$ (b), the modeling of other backgrounds (c) and instrumental (d) systematic uncertainties for full fit to data.

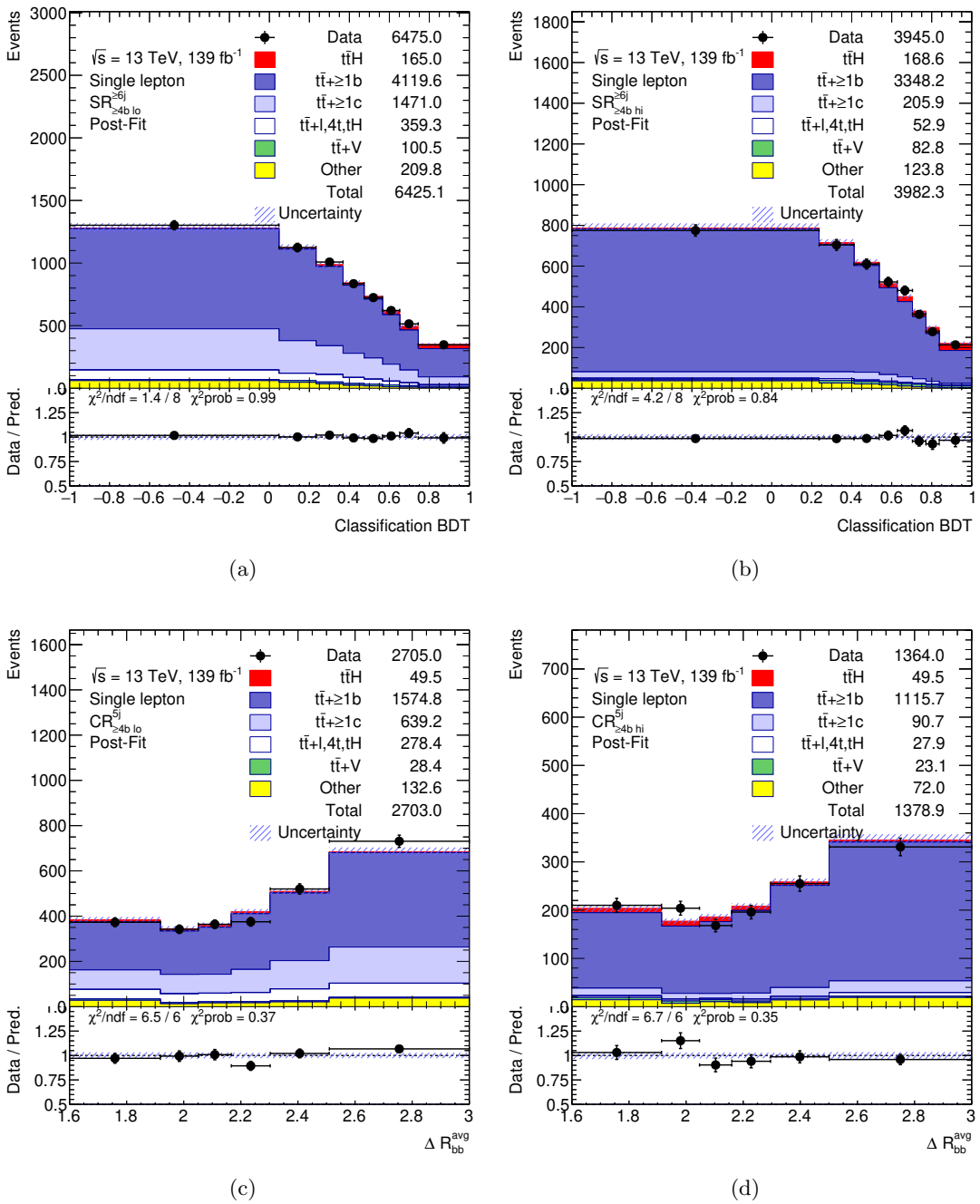
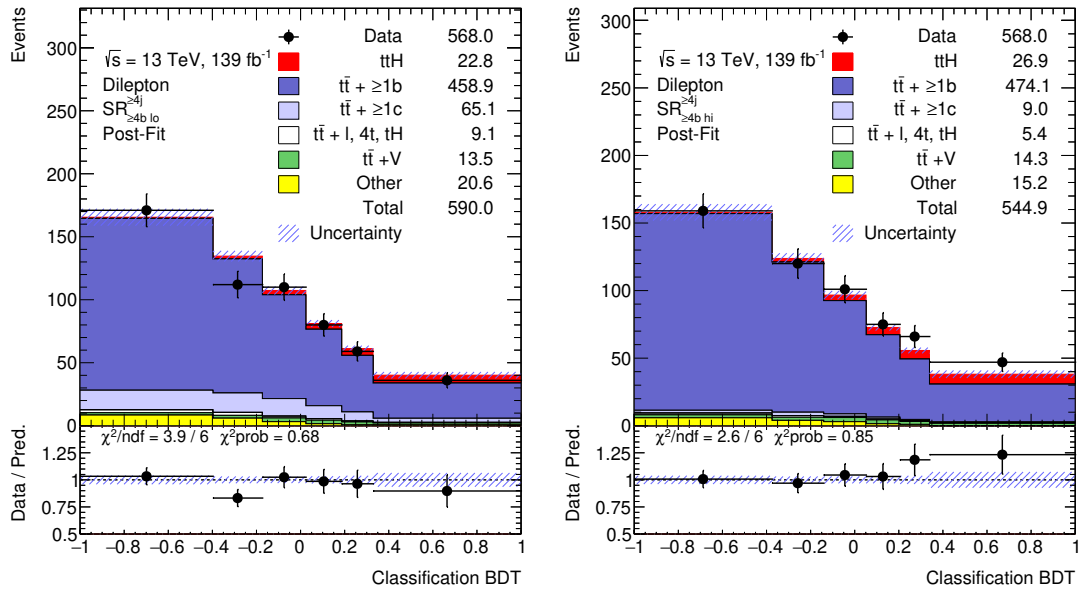


Figure 8.15: Post-fit modeling of the four single lepton analysis regions, (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$, (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$, (c) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{lo}}$ and (d) $\mathrm{SR}^{5j}_{\geq 4b\ \mathrm{hi}}$, after performing the full signal+background fit, displayed as a function of the variable used, the $\Delta R^{\mathrm{avg}}_{bb}$ for the 5 jet regions and the classification BDT for 6 jet regions.



(a) (b)

Figure 8.16: Post-fit modeling of the two dilepton signal regions, (a) $\mathrm{SR}^{\geq 4j}_{\geq 4b\ \mathrm{lo}}$ and (b) $\mathrm{SR}^{\geq 4j}_{\geq 4b\ \mathrm{hi}}$, after performing the full signal+background fit, displayed as a function of the classification BDT.

Impact of systematic uncertainties

The ranking plot of the combined fit is shown in figure 8.17, displaying features similar to the results of the Asimov expectation (see figure 8.7). The impact on the $\mu_{t\bar{t}H}$ is dominated by the $t\bar{t}+\geq 1b$ NLO match systematic, followed closely by the $t\bar{t}+\geq 1b$ ISR. The fact that they are both significantly shifted compared to their nominal value and have a large constraint demonstrates their importance to the background modeling. This was already clear for the ISR because of its impact on the jet multiplicity. This pull could be avoided in the future by shifting of the renormalization and factorization scale in the nominal sample.

The large variation of the $t\bar{t}+\geq 1b$ NLO match points to a tension in the model which cannot be easily interpreted for two reasons. First, the systematic variation is defined between two $t\bar{t}$+jets models and as such is probably overestimated and does not precisely describe NLO matching of the $t\bar{t}b\bar{b}$ in the matrix element. Second, the systematic uncertainty is defined by a comparison of two models and the negative value of the nuisance parameter does not directly correspond to a specific model.

The next few systematics in the ranking, coming mainly from $t\bar{t}+\geq 1b$ and $t\bar{t}H$ modeling, have a similar impact and display only a small change in their order with respect to the Asimov fit. Neither of them is significantly pulled, meaning their actual importance to the modeling is not as large. The first experimental systematic is the first eigenvalue of the efficiency of b-jets in the b-tagging (bTag b-jets EV0) in the seventh place.

The impact of the systematics was also assessed according to their contribution to the total uncertainty. The estimation is done by fixing given group of systematics to their optimal value and performing the fit again. The difference between the nominal fit and this variation (subtracted in quadrature) estimates the contribution of the given group to the error of the $t\bar{t}H$. This was performed in groups of parameters instead of individually. The

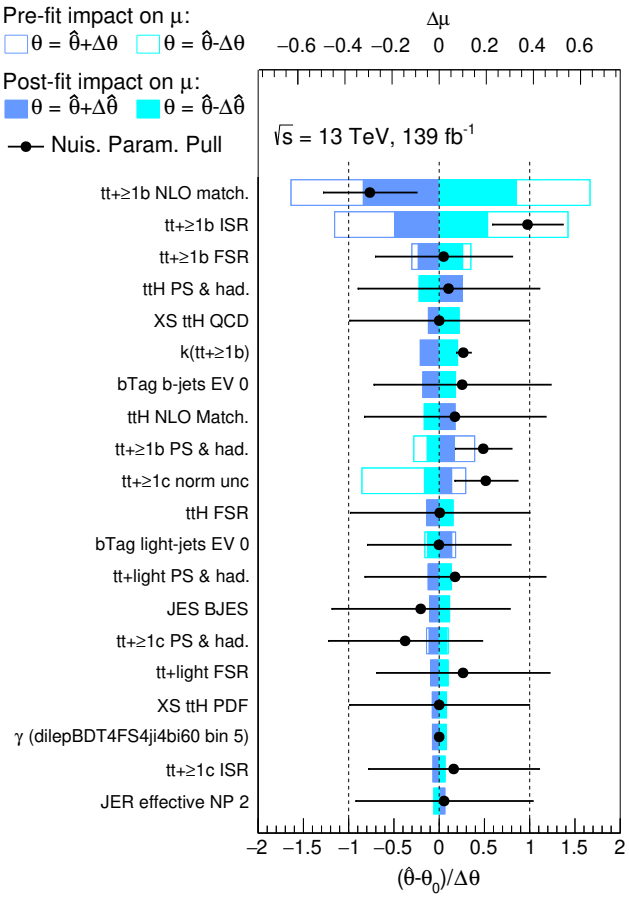


Figure 8.17: Ranking plot for 20 nuisance parameters with the largest impact on the parameter of interest $\mu_{t\bar{t}H}$ in the full fit to the data.. It is shown for the pre-fit uncertainty by the empty box and post-fit by the filled box. The post-fit shifts and constraints of the systematics are displayed as the black markers and the horizontal line respectively.

Group	$+\Delta\mu_{t\bar{t}H}$	$-\Delta\mu_{t\bar{t}H}$
$t\bar{t}H$ syst.	+0.215	-0.085
$t\bar{t}+\geq 1b$ norm.	+0.077	-0.095
$t\bar{t}+\geq 1b$ NLO match	+0.343	-0.316
$t\bar{t}+\geq 1b$ ISR	+0.211	-0.190
other $t\bar{t}+\geq 1b$ syst.	+0.108	-0.110
other $t\bar{t}$+jets	+0.102	-0.085
other bkg.	+0.037	-0.030
Experimental	+0.151	-0.113
MC stat. uncertainty	+0.049	-0.051
Total	+0.469	-0.416

Table 8.3: Contribution of various group of systematic uncertainties on the measured $\mu_{t\bar{t}H}$ uncertainty.

two dominant systematic $t\bar{t}+\geq 1b$ NLO match and $t\bar{t}+\geq 1b$ ISR are still shown individually, but the two remaining $t\bar{t}+\geq 1b$ systematics are grouped under *other* $t\bar{t}+\geq 1b$ label. The $t\bar{t}H$ systematics are grouped under $t\bar{t}H$ label and the other $t\bar{t}$+jets systematics are under other $t\bar{t}$+jets. Finally, modeling systematics of the small backgrounds are grouped under *other bkg.* The remaining two groups are the experimental systematics and the grouped impact of Monte Carlo statistics of the nominal samples.

The results can be found in table 8.3. It generally confirms what was already apparent from the ranking plot. Interestingly, combined impact of the $t\bar{t}H$ systematic variations has a much larger impact on the positive variation of the $\mu_{t\bar{t}H}$ than the negative. This was already the case in the previous iteration of the analysis[8].

Signal performance in individual channels

The two analysis channels give quite different values of the $\mu_{t\bar{t}H}$ when fitted independently. Its value in the single lepton channel is lower than the SM prediction with $\mu_{t\bar{t}H}^{\text{single lepton}} = 0.54^{+0.61}_{-0.58}$ while in the dilepton channel it is higher with $\mu_{t\bar{t}H}^{\text{dilepton}} = 1.43^{+0.69}_{-0.62}$. The combination then reconvenes the two channels at value of $\mu_{t\bar{t}H} = 0.84^{+0.45}_{-0.39}(\text{syst.}) \pm 0.21(\text{stat.})$.

In order to estimate how the signal in the individual channels would perform if the background were constrained by both, the value of the $\mu_{t\bar{t}H}$ normalization is decorrelated across the two channels and the combined fit is repeated. Such fit results in $\mu_{t\bar{t}H}^{\text{single lepton}} = 0.51^{+0.53}_{-0.50}$ and $\mu_{t\bar{t}H}^{\text{dilepton}} = 1.20^{+0.59}_{-0.53}$. While the former does not change much compared to the fit in a single channel, the dilepton signal gets smaller by approximately 20%. This suggests, that even though the values of modeling nuisance parameters differ quite a lot between the single lepton and combined measurement, as discussed in section 8.3, the impact of the different measured NPs on the measured value of $\mu_{t\bar{t}H}$ in the single lepton channel is negligible.

A comparison of the nuisance parameters between the fits with a single and decorrelated $\mu_{t\bar{t}H}$ can be found in figure 8.18. It shows that adding additional degree of freedom on the signal does not change the background modeling.

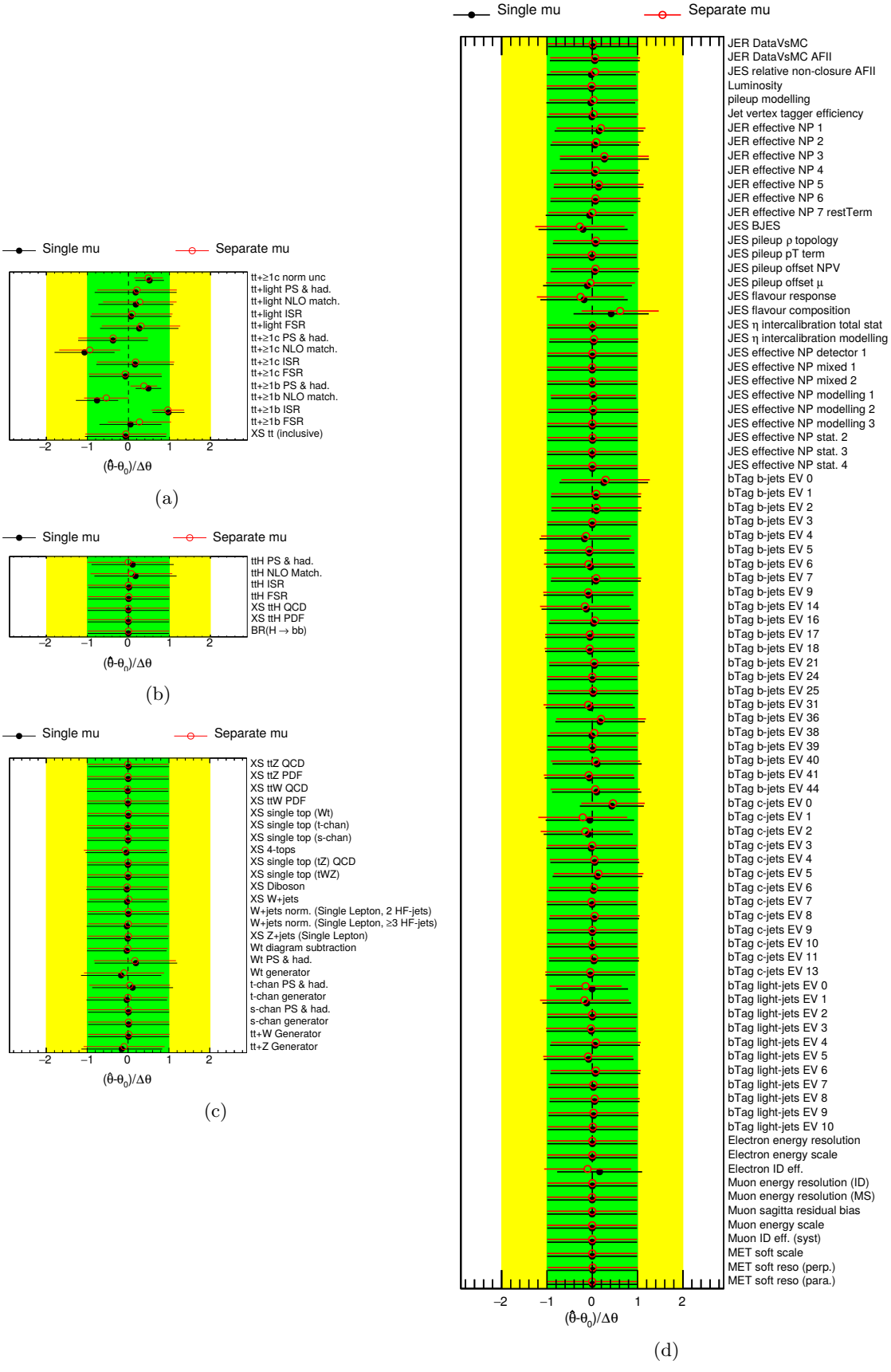


Figure 8.18: Resulting pulls and constraints of the $t\bar{t}$+jets modeling (a), the modeling the $t\bar{t}H$ (b) , the modeling of other backgrounds (c) and instrumental (d) systematic uncertainties for full fit to data with a single $\mu_{t\bar{t}H}$ and one decorrelated across the two analysis channels.

8.6 Interpretation of the results

The analysis reports a signal strength $\mu_{t\bar{t}H} = 0.84^{+0.45}_{-0.39}(\text{syst.}) \pm 0.21(\text{stat.})$, which can be directly compared to the result of the previous measurement $\mu_{t\bar{t}H} = 0.84^{+0.57}_{-0.54}(\text{syst.}) \pm 0.29(\text{stat.})$[8]. The statistical uncertainties are reduced from 29% to 21%, which is a smaller effect than expected out of the four-fold increase in statistics. This is probably caused by the tighter selection introduced to simplify the analysis. The systematic uncertainties are reduced to a similar degree by approximately 10%.

The new analysis has a 2.3σ expected significance, compared to the 1.6σ of the previous ATLAS mesasurement. The observed significance is then slightly lower in both cases: 1.9σ for the newer measurement and 1.4σ for the previous measurement.

The main limitation of the analysis is the modeling of the $t\bar{t}b\bar{b}$ background. The $t\bar{t}+\geq 1b$ ISR systematic variation is understood, as it improves the modeling of the jet multiplicity by introducing lower renormalization and factorization scales in the matrix element. Currently, the systematic has to be pulled by approximately 1σ from its central value. Producing a new nominal sample with such scale would be preferable to a large pull which has a penalty in the likelihood.

The interpretation of the $t\bar{t}+\geq 1b$ NLO match systematic is not so straightforward, though it points at a possible mis-modeling of the matching of the NLO matrix element to the parton shower. Furthermore, the systematic is determined in the $t\bar{t}$+jets sample, which does not accurately describe the $t\bar{t}b\bar{b}$ process. Producing proper variation of the NLO matching, and generally having more alternatives to the $t\bar{t}b\bar{b}$ with the b quarks included in the matrix element, could be used to assess this effect in more detail, however, at the time of writing of this thesis no alternatives with sufficient statistics was available.

Though the values of the $\mu_{t\bar{t}H}$ when decorrelated between the single lepton and dilepton channel show a discrepancy, it is smaller than what was reported in the previous measurement, where they were $\mu_{t\bar{t}H}^{\text{single lepton}} = 0.95^{+0.65}_{-0.65}$ and $\mu_{t\bar{t}H}^{\text{dilepton}} = -0.24^{+1.02}_{-1.05}$. The difference between the two values is larger but it is also covered by bigger uncertainties (especially in the dilepton channel). Interestingly, the discrepancy between the two channel has the opposite direction. Currently, the reason is not understood, since it does not seem to come only from the increase in statistics. The detector reconstruction changed significantly between the two measurements, making the investigation difficult.

Nevertheless, the results of the two measurements are compatible and the new measurement presents an improvement compared to the previous result, with a measured significance of 1.9σ.

CHAPTER 9

Conclusions and outlook

The coupling of the Higgs boson to the heaviest quark, the top quark, was until recently known only through indirect measurements[4]. The recent analysis of the Higgs production associated with pair of top quarks ($t\bar{t}H$)[5, 6] provided a first estimation of this coupling in a combined measurement of several channels, defined by the decay products of the Higgs boson. The result showed an agreement with the Standard Model prediction, however, the uncertainty is still quite large and more precise measurement can find a discrepancy. Furthermore, a differential measurement would probe the CP properties of the coupling, which are not yet established. Both aspects provide a strong motivation to improve and advance the $t\bar{t}H$ measurement.

One of the channels where the $t\bar{t}H$ cross-section was measured is a final state where the Higgs decays into a pair of b quarks. Though it has the highest branching ratio, the precision of the measurement in this channel is strongly limited by a low precision of the background modeling.

The background is dominated by $t\bar{t}b\bar{b}$, a $t\bar{t}$ process with additional b quarks introduced into the event from a splitting of radiated gluon. A large number of heavy flavor jets in the final state is currently not well modeled. Numerous systematic uncertainties were accounted for, decreasing the sensitivity of the measurement.

In order to maximize the separation of the signal from the background, a multi-variate algorithm is employed in the signal rich regions. In addition, various control regions are used to better constrain modeling of the backgrounds. This dissertation presents in detail an updated measurement of the $t\bar{t}H(b\bar{b})$ at 13 TeV in the single lepton channel, reporting a signal strength $\mu_{t\bar{t}H}^{\text{single lepton}} = 0.54^{+0.61}_{-0.58}$, a value compatible with the Standard Model prediction.

In the previous iteration of the analysis[8], low statistics of the Monte Carlo samples had a large impact on the result. In the analysis presented in this thesis, the statistics was significantly increased and its effects on the measurement were assessed in detail through a bootstrap method. The impact on the results of the fit was found to be negligible.

The single lepton channel was also combined with a measurement in the dilepton channel. The combined analysis reports a measured signal strength $\mu_{t\bar{t}H} = 0.84^{+0.45}_{-0.39}(\text{syst.}) \pm 0.21(\text{stat.})$, providing a $t\bar{t}H(b\bar{b})$ measurement with significance 1.9σ (2.3σ expected), a small improvement compared to the previous iteration with 1.4σ (1.6σ expected).

One of the difficulties of the $t\bar{t}H(b\bar{b})$ measurement is its inaccurate modeling of $t\bar{t}+\geq 1c$ and $t\bar{t}+\geq 1b$ fractions in the $t\bar{t}$+jets process. This analysis reports approximately 1.6 times higher contribution of the former and 1.3 times larger of the latter compared to the

MC expectation, a result in agreement with the previous $t\bar{t}H(b\bar{b})$ measurement[8] and a dedicated analysis of the $t\bar{t}b\bar{b}$ process[122].

The $t\bar{t}H(b\bar{b})$ analysis is still limited by systematic uncertainties on the $t\bar{t}b\bar{b}$ final state. One of the limitations could be mitigated by lowering of the renormalization and factorization scale of the nominal $t\bar{t}b\bar{b}$ sample, as explained in section 8.3.1. There is, however, still a remaining mis-modeling covered by two point systematic uncertainties which are not so easily interpreted. Detailed study of alternative models, possibly in a dedicated $t\bar{t}b\bar{b}$ measurement, is needed to improve the modeling of the $t\bar{t}b\bar{b}$ sample. Better understanding of the background would significantly improve the sensitivity of the $t\bar{t}H(b\bar{b})$ measurement.

However, improved $t\bar{t}b\bar{b}$ estimation requires good Monte Carlo generators to properly define the model, instead of using extrapolation from the inclusive $t\bar{t}$ sample as used in this analysis. This for example means alternatives to the NLO matching and to the PS and hadronization models of the nominal $t\bar{t}b\bar{b}$ sample. Good candidates are other generators using the same $t\bar{t}b\bar{b}$ matrix element but alternative models for the other parts of the event generation.

Furthermore, including matrix elements with different numbers of jets in the final state can lead to a better description of the jet kinematic properties, while dedicated generators with massive b-quarks can be used to improve the description of b-jets. To provide a full picture, these matrix elements need to be properly merged, which can be done with dedicated algorithms[148–150] to merge NLO matrix elements or methods to properly merge matrix elements with and without massive quarks[151].

The future analysis can also take advantage of the newest reconstruction techniques to improve the background rejection. This means for example using the PFlow jets[99] and DL1r b-tagging algorithms[106], described previously in chapter 5.

Bibliography

[1] ATLAS Collaboration, *Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*, Physics Letters B **716** (2012) 1, ISSN: 0370-2693, URL: http://www.sciencedirect.com/science/article/pii/S037026931200857X (cit. on pp. 1, 11).

[2] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Physics Letters B **716** (2012) 30, ISSN: 0370-2693, URL: http://www.sciencedirect.com/science/article/pii/S0370269312008581 (cit. on pp. 1, 11).

[3] D. de Florian, C. Grojean, F. Maltoni, C. Mariotti, and A. Nikitenko, *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*, 2016, arXiv: 1610.07922 [hep-ph] (cit. on pp. 1, 11, 12, 73).

[4] ATLAS Collaboration, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s} = 7$ and 8 TeV*, Journal of High Energy Physics **2016** (2016), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP08(2016)045 (cit. on pp. 1, 127).

[5] CMS Collaboration, *Observation of ttH Production*, Physical Review Letters **120** (2018), ISSN: 1079-7114, URL: http://dx.doi.org/10.1103/PhysRevLett.120.231801 (cit. on pp. 1, 13, 127).

[6] ATLAS Collaboration, *Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector*, Physics Letters B **784** (2018) 173, ISSN: 0370-2693, URL: http://dx.doi.org/10.1016/j.physletb.2018.07.035 (cit. on pp. 1, 13, 127).

[7] M. Tanabashi, K. Hagiwara, K. Hikasa, and Nakamura, *Particle Data Group*, Phys. Rev. D **98** (3 2018 and 2019 update) 030001, URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001 (cit. on p. 1).

[8] ATLAS Collaboration, *Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. **D97** (2018) 072016, arXiv: 1712.08895 [hep-ex] (cit. on pp. 2, 54, 55, 62, 73, 74, 89, 124, 126–128, 150).

[9] J. Schwinger, ed., *Selected Papers on Quantum Electrodynamics*, Dover Publications, 1958, ISBN: 978-0-486-60444-2, 978-0-486-60444-2 (cit. on pp. 3, 5).

[10] S. L. Glashow, *Partial-symmetries of weak interactions*, Nuclear Physics **22** (1961) 579, ISSN: 0029-5582, URL: http://www.sciencedirect.com/science/article/pii/0029558261904692 (cit. on p. 3).

[11] S. Weinberg, *A Model of Leptons*, Phys. Rev. Lett. **19** (21 1967) 1264, URL: https://link.aps.org/doi/10.1103/PhysRevLett.19.1264 (cit. on p. 3).

[12] A. Salam, *Weak and Electromagnetic Interactions*, Conf. Proc. C **680519** (1968) 367 (cit. on p. 3).

[13] G. 't Hooft and M. Veltman, *Regularization and renormalization of gauge fields*, Nuclear Physics B **44** (1972) 189, ISSN: 0550-3213, URL: http://www.sciencedirect.com/science/article/pii/0550321372902799 (cit. on p. 3).

[14] F. Englert and R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321, ed. by J. Taylor (cit. on p. 4).

[15] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (16 1964) 508, URL: https://link.aps.org/doi/10.1103/PhysRevLett.13.508 (cit. on p. 4).

[16] P. W. Higgs, *Spontaneous Symmetry Breakdown without Massless Bosons*, Phys. Rev. **145** (1966) 1156 (cit. on p. 4).

[17] G. Serra, *Standard Model*, Accessed: 2016-04-07, URL: http://www.physik.uzh.ch/groups/serra/StandardModel.html (cit. on p. 4).

[18] J. Horejsi, *Fundamentals of electroweak theory*, 2002 (cit. on pp. 5, 7).

[19] G. Arnison et al., *Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider*, Phys. Lett. B **126** (1983) 398 (cit. on p. 7).

[20] P. Bagnaia et al., *Evidence for $Z^0 \to e^+e^-$ at the CERN $\bar{p}p$ Collider*, Phys. Lett. B **129** (1983) 130 (cit. on p. 7).

[21] M. Tanabashi et al., *Review of Particle Physics*, Phys. Rev. D **98** (3 2018) 030001, URL: https://link.aps.org/doi/10.1103/PhysRevD.98.030001 (cit. on pp. 9–11, 14, 16, 17, 19–21, 45, 78, 79).

[22] P. Skands, *Introduction to QCD*, Searching for New Physics at Small and Large Scales (2013), URL: http://dx.doi.org/10.1142/9789814525220_0008 (cit. on pp. 9, 10).

[23] M. L. Perl et al., *Evidence for Anomalous Lepton Production in $e^+ - e^-$ Annihilation*, Phys. Rev. Lett. **35** (22 1975) 1489, URL: https://link.aps.org/doi/10.1103/PhysRevLett.35.1489 (cit. on p. 10).

[24] S. W. Herb et al., *Observation of a Dimuon Resonance at 9.5 GeV in 400-GeV Proton-Nucleus Collisions*, Phys. Rev. Lett. **39** (5 1977) 252, URL: https://link.aps.org/doi/10.1103/PhysRevLett.39.252 (cit. on p. 10).

[25] K. Kodama et al., *Observation of tau neutrino interactions*, Physics Letters B **504** (2001) 218, ISSN: 0370-2693, URL: http://dx.doi.org/10.1016/S0370-2693(01)00307-0 (cit. on p. 10).

[26] F. Abe et al., *Observation of Top Quark Production in pp Collisions with the Collider Detector at Fermilab*, Physical Review Letters **74** (1995) 2626, ISSN: 1079-7114, URL: http://dx.doi.org/10.1103/PhysRevLett.74.2626 (cit. on p. 11).

[27] S. Abachi et al., *Observation of the Top Quark*, Physical Review Letters **74** (1995) 2632, ISSN: 1079-7114, URL: http://dx.doi.org/10.1103/PhysRevLett.74.2632 (cit. on p. 11).

[28] D. Binosi and L. Theußl, *JaxoDraw: A graphical user interface for drawing Feynman diagrams*, Computer Physics Communications **161** (2004) 76, ISSN: 0010-4655, URL: http://www.sciencedirect.com/science/article/pii/S0010465504002115 (cit. on pp. 11, 12, 17, 18, 63).

[29] ATLAS Collaboration, *Measurements of Higgs boson properties in the diphoton decay channel with 36 fb^{-1} of pp collision data at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **98** (2018) 052005, arXiv: 1802.04146 [hep-ex] (cit. on p. 13).

[30] ATLAS Collaboration, *Measurement of the Higgs boson coupling properties in the $H \to ZZ^* \to 4\ell$ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector*, JHEP **03** (2018) 095, arXiv: 1712.02304 [hep-ex] (cit. on p. 13).

[31] ATLAS Collaboration, *Evidence for the associated production of the Higgs boson and a top quark pair with the ATLAS detector*, Phys. Rev. D **97** (2018) 072003, arXiv: 1712.08891 [hep-ex] (cit. on pp. 13, 64).

[32] ATLAS Collaboration, *Analysis of $t\bar{t}H$ and $t\bar{t}W$ production in multilepton final states with the ATLAS detector*, tech. rep. ATLAS-CONF-2019-045, CERN, 2019, URL: http://cds.cern.ch/record/2693930 (cit. on pp. 13, 64).

[33] Y. L. Dokshitzer, *Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics.*, Sov. Phys. JETP **46** (1977) 641 (cit. on p. 15).

[34] V. Gribov and L. Lipatov, *e+ e–pair annihilation and deep inelastic ep-scattering in perturbation theory*, Yadern. Fiz **15** (1972) 1218 (cit. on p. 15).

[35] G. Altarelli and G. Parisi, *Asymptotic freedom in parton language*, Nuclear Physics B **126** (1977) 298 (cit. on p. 15).

[36] J. Rojo et al., *The PDF4LHC report on PDFs and LHC data: results from Run I and preparation for Run II*, Journal of Physics G: Nuclear and Particle Physics **42** (2015) 103103, URL: https://doi.org/10.1088%2F0954-3899%2F42%2F10%2F103103 (cit. on pp. 15, 16).

[37] ATLAS Collaboration, *The ATLAS Experiment at the CERN Large Hadron Collider*, Journal of Instrumentation **3** (2008) S08003, URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08003 (cit. on pp. 15, 23, 27–32, 34–37).

[38] T. Sjostrand, S. Mrenna, and P. Skands, *A brief introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852, arXiv: 0710.3820 [hep-ph] (cit. on pp. 17, 22).

[39] G. Marchesini and B. Webber, *Monte Carlo Simulation of General Hard Processes with Coherent QCD Radiation*, Nucl. Phys. B **310** (1988) 461 (cit. on p. 17).

[40] J. Bellm et al., *Herwig 7.0/Herwig++ 3.0 release note*, Eur. Phys. J. C **76** (2016) 196, arXiv: 1512.01178 [hep-ph] (cit. on pp. 17, 22, 56, 57, 60).

[41] T. Sjostrand, *A Model for Initial State Parton Showers*, Phys. Lett. B **157** (1985) 321 (cit. on p. 19).

[42] G. Gustafson and U. Pettersson, *Dipole formulation of QCD cascades*, Nuclear Physics B **306** (1988) 746, ISSN: 0550-3213, URL: http://www.sciencedirect.com/science/article/pii/0550321388904415 (cit. on p. 19).

[43] E. Bothmann et al., *Event Generation with Sherpa 2.2*, (2019), arXiv: 1905.09127 [hep-ph] (cit. on pp. 19, 22, 61).

[44] E. Norrbin and T. Sjöstrand, *QCD radiation off heavy particles*, Nuclear Physics B **603** (2001) 297, ISSN: 0550-3213, URL: http://dx.doi.org/10.1016/S0550-3213(01)00099-2 (cit. on p. 19).

[45] S. Catani, S. Dittmaier, M. H. Seymour, and Z. Trócsányi, *The dipole formalism for next-to-leading order QCD calculations with massive partons*, Nuclear Physics B **627** (2002) 189, ISSN: 0550-3213, URL: http://dx.doi.org/10.1016/S0550-3213(02)00098-6 (cit. on p. 19).

[46] S. Frixione, P. Nason, and C. Oleari, *Matching NLO QCD computations with Parton Shower simulations: the POWHEG method*, JHEP **11** (2007) 070, arXiv: 0709.2092 [hep-ph] (cit. on pp. 20, 22, 59, 61).

[47] R. Frederix and S. Frixione, *Merging meets matching in MC@NLO*, JHEP **12** (2012) 061, arXiv: 1209.6215 [hep-ph] (cit. on pp. 20, 22, 61).

[48] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjostrand, *Parton Fragmentation and String Dynamics*, Phys. Rept. **97** (1983) 31 (cit. on p. 20).

[49] B. Andersson, *The Lund model*, vol. 7, Cambridge University Press, 2005, ISBN: 978-0-521-01734-3, 978-0-521-42094-5, 978-0-511-88149-7 (cit. on p. 20).

[50] R. D. Field and S. Wolfram, *A QCD Model for e+ e- Annihilation*, Nucl. Phys. B **213** (1983) 65 (cit. on p. 21).

[51] D. Amati and G. Veneziano, *Preconfinement as a Property of Perturbative QCD*, Phys. Lett. B **83** (1979) 87 (cit. on p. 21).

[52] T. Sjöstrand and M. van Zijl, *A multiple-interaction model for the event structure in hadron collisions*, Phys. Rev. D **36** (7 1987) 2019, URL: https://link.aps.org/doi/10.1103/PhysRevD.36.2019 (cit. on p. 21).

[53] S. Alioli, P. Nason, C. Oleari, and E. Re, *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043, arXiv: 1002.2581 [hep-ph] (cit. on pp. 22, 59, 61).

[54] L. Evans and P. Bryant, *LHC Machine*, Journal of Instrumentation **3** (2008) S08001, URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08001 (cit. on pp. 23, 25).

[55] *About CERN*, Accessed: 2020-05-25, URL: https://home.cern/about (cit. on p. 23).

[56] *CERN's accelerators*, Accessed: 2020-05-26, URL: https://home.cern/science/accelerators/ (cit. on p. 23).

[57] E. Mobs, *The CERN accelerator complex. Complexe des accélérateurs du CERN*, (2016), General Photo, Accessed: 2020-05-25, URL: https://cds.cern.ch/record/2197559 (cit. on p. 24).

[58] ALICE Collaboration, *The ALICE experiment at the CERN LHC*, Journal of Instrumentation **3** (2008) S08002, URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08002 (cit. on p. 23).

[59] LHCb Collaboration, *The LHCb Detector at the LHC*, Journal of Instrumentation **3** (2008) S08005, URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08005 (cit. on p. 25).

[60] CMS Collaboration, *The CMS experiment at the CERN LHC*, Journal of Instrumentation **3** (2008) S08004, URL: https://doi.org/10.1088%2F1748-0221%2F3%2F08%2Fs08004 (cit. on p. 25).

[61] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC*, tech. rep. ATLAS-CONF-2019-021, CERN, 2019, URL: http://cds.cern.ch/record/2677054 (cit. on pp. 25, 26, 36, 73).

[62] S. van der Meer, *Calibration of the effective beam height in the ISR*, tech. rep. CERN-ISR-PO-68-31. ISR-PO-68-31, CERN, 1968, URL: https://cds.cern.ch/record/296752 (cit. on p. 25).

[63] ATLAS Collaboration, *Luminosity determination in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector at the LHC*, The European Physical Journal C **76** (2016), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-016-4466-1 (cit. on pp. 25, 36).

[64] *ATLAS luminosity public results*, Accessed: 2020-02-26, URL: https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2 (cit. on p. 26).

[65] L. de Nooij, *The $\Phi(1020)$-meson production cross section measured with the ATLAS detector at $\sqrt{S} = 7$ TeV*, (2014), URL: https://cds.cern.ch/record/1701359 (cit. on p. 29).

[66] ATLAS Collaboration, *ATLAS Insertable B-Layer Technical Design Report*, tech. rep. CERN-LHCC-2010-013. ATLAS-TDR-19, 2010, URL: https://cds.cern.ch/record/1291633 (cit. on p. 30).

[67] ATLAS Collaboration, *IBL Efficiency and Single Point Resolution in Collision Events*, tech. rep. ATL-INDET-PUB-2016-001, CERN, 2016, URL: https://cds.cern.ch/record/2203893 (cit. on p. 30).

[68] ATLAS Collaboration, *Alignment of the ATLAS Inner Detector and its Performance in 2012*, tech. rep. ATLAS-CONF-2014-047, CERN, 2014, URL: https://cds.cern.ch/record/1741021 (cit. on pp. 30, 31).

[69] V. Tisserand, *Optimization of the ATLAS detector for the search of Higgs boson decaying into two photons at the LHC*, Theses, 1997, URL: https://tel.archives-ouvertes.fr/tel-00544731 (cit. on p. 33).

[70] ATLAS Collaboration, *Observation and measurement of forward proton scattering in association with lepton pairs produced via the photon fusion mechanism at ATLAS*, tech. rep. ATLAS-CONF-2020-041, CERN, 2020, URL: http://cds.cern.ch/record/2727863 (cit. on p. 36).

[71] P. Jenni and M. Nessi, *ATLAS Forward Detectors for Luminosity Measurement and Monitoring*, tech. rep. CERN-LHCC-2004-010. LHCC-I-014, revised version number 1 submitted on 2004-03-22 14:56:11: CERN, 2004, URL: http://cds.cern.ch/record/721908 (cit. on p. 36).

[72] S. A. Khalek et al., *The ALFA Roman Pot detectors of ATLAS*, Journal of Instrumentation **11** (2016) P11013, ISSN: 1748-0221, URL: http://dx.doi.org/10.1088/1748-0221/11/11/P11013 (cit. on pp. 36, 37).

[73] P. Jenni, M. Nessi, and M. Nordberg, *Zero Degree Calorimeters for ATLAS*, tech. rep. CERN-LHCC-2007-001. LHCC-I-016, CERN, 2007, URL: http://cds.cern.ch/record/1009649 (cit. on p. 36).

[74] L. Adamczyk et al., *Technical Design Report for the ATLAS Forward Proton Detector*, tech. rep. CERN-LHCC-2015-009. ATLAS-TDR-024, 2015, URL: https://cds.cern.ch/record/2017378 (cit. on pp. 36, 37).

[75] J. Lange et al., *Beam tests of an integrated prototype of the ATLAS Forward Proton detector*, Journal of Instrumentation **11** (2016) P09005, URL: https://doi.org/10.1088%2F1748-0221%2F11%2F09%2Fp09005 (cit. on p. 37).

[76] ATLAS Collaboration, *Operation of the ATLAS trigger system in Run 2*, (2020), arXiv: 2007.12539 [physics.ins-det] (cit. on p. 37).

[77] S. Artz et al., *Upgrade of the ATLAS Central Trigger for LHC Run-2*, Journal of Instrumentation **10** (2015) C02030, URL: https://doi.org/10.1088%2F1748-0221%2F10%2F02%2Fc02030 (cit. on p. 37).

[78] ATLAS Collaboration, *Performance of the ATLAS trigger system in 2015*, The European Physical Journal C **77** (2017), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-017-4852-3 (cit. on p. 37).

[79] ATLAS Collaboration, *The ATLAS Simulation Infrastructure*, The European Physical Journal C **70** (2010) 823, ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-010-1429-9 (cit. on p. 38).

[80] S. Agostinelli et al., *Geant4—a simulation toolkit*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **506** (2003) 250, ISSN: 0168-9002, URL: http://www.sciencedirect.com/science/article/pii/S0168900203013688 (cit. on p. 38).

[81] E. Richter-Was, D. Froidevaux, and L. Poggioli, *ATLFAST 2.0 a fast simulation package for ATLAS*, tech. rep. ATL-PHYS-98-131, CERN, 1998, URL: https://cds.cern.ch/record/683751 (cit. on p. 38).

[82] A. Held, *Search for the production of Higgs bosons in association with top quarks and decaying into bottom quark pairs with the ATLAS detector*, (2019), URL: https://cds.cern.ch/record/2693579 (cit. on p. 40).

[83] J. Pequenao and P. Schaffner, "How ATLAS detects particles: diagram of particle paths in the detector", 2013, URL: https://cds.cern.ch/record/1505342 (cit. on pp. 40, 41).

[84] T. Cornelissen et al., *Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)*, tech. rep. ATL-SOFT-PUB-2007-007. ATL-COM-SOFT-2007-002, CERN, 2007, URL: https://cds.cern.ch/record/1020106 (cit. on pp. 40, 41).

[85] ATLAS Collaboration, *Performance of the ATLAS Track Reconstruction Algorithms in Dense Environments in LHC Run 2*, Eur. Phys. J. C **77** (2017) 673, arXiv: 1704.07983 [hep-ex] (cit. on p. 41).

[86] R. Frühwirth, *Application of Kalman filtering to track and vertex fitting*, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **262** (1987) 444, ISSN: 0168-9002, URL: http://www.sciencedirect.com/science/article/pii/0168900287908874 (cit. on p. 41).

[87] ATLAS Collaboration, *Reconstruction of primary vertices at the ATLAS experiment in Run 1 proton–proton collisions at the LHC*, Eur. Phys. J. C **77** (2017) 332, arXiv: 1611.10235 [physics.ins-det] (cit. on p. 42).

[88] ATLAS Collaboration, *Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s}$ = 13 TeV*, Eur. Phys. J. C **79** (2019) 639, arXiv: 1902.04655 [physics.ins-det] (cit. on pp. 42–45).

[89] ATLAS Collaboration, *Electron and photon reconstruction and identification in ATLAS: expected performance at high energy and results at 900 GeV*, tech. rep. ATLAS-CONF-2010-005, CERN, 2010, URL: https://cds.cern.ch/record/1273197 (cit. on p. 43).

[90] P. Sommer, Private Communication, 2020 (cit. on p. 43).

[91] ATLAS Collaboration, *Electron efficiency measurements with the ATLAS detector using 2012 LHC proton–proton collision data*, The European Physical Journal C **77** (2017), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-017-4756-2 (cit. on p. 45).

[92] ATLAS Collaboration, *Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ =13 TeV*, Eur. Phys. J. C **76** (2016) 292, arXiv: 1603.05598 [hep-ex] (cit. on pp. 45, 46).

[93] G. P. Salam, *Towards Jetography*, 2009, arXiv: 0906.1833 [hep-ph] (cit. on p. 47).

[94] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-ktjet clustering algorithm*, Journal of High Energy Physics **2008** (2008) 063, URL: https://doi.org/10.1088%2F1126-6708%2F2008%2F04%2F063 (cit. on pp. 47, 48).

[95] M. Cacciari, G. P. Salam, and G. Soyez, *FastJet User Manual*, Eur. Phys. J. C **72** (2012) 1896, arXiv: 1111.6097 [hep-ph] (cit. on p. 48).

[96] ATLAS Collaboration, *Jet energy scale measurements and their systematic uncertainties in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, Phys. Rev. D **96** (2017) 072002, arXiv: 1703.09665 [hep-ex] (cit. on pp. 48, 49).

[97] ATLAS Collaboration, *Determination of jet calibration and energy resolution in proton-proton collisions at $\sqrt{s}$ = 8 TeV using the ATLAS detector*, (2019), arXiv: 1910.04482 [hep-ex] (cit. on p. 49).

[98] ATLAS Collaboration, *Performance of pile-up mitigation techniques for jets in pp collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector*, Eur. Phys. J. C **76** (2016) 581, arXiv: 1510.03823 [hep-ex] (cit. on p. 49).

[99] ATLAS Collaboration, *Jet reconstruction and performance using particle flow with the ATLAS Detector*, Eur. Phys. J. C **77** (2017) 466, arXiv: 1703.10485 [hep-ex] (cit. on pp. 49, 128).

[100] ATLAS Collaboration, *ATLAS b-jet identification performance and efficiency measurement with $t\bar{t}$ events in pp collisions at $\sqrt{s} = 13$ TeV*, Eur. Phys. J. C **79** (2019) 970, arXiv: 1907.05120 [hep-ex] (cit. on p. 50).

[101] ATLAS Collaboration, *Optimisation of the ATLAS b-tagging performance for the 2016 LHC Run*, tech. rep. ATL-PHYS-PUB-2016-012, CERN, 2016, URL: https://cds.cern.ch/record/2160731 (cit. on pp. 50, 52).

[102] ATLAS Collaboration, *Secondary vertex finding for jet flavour identification with the ATLAS detector*, tech. rep. ATL-PHYS-PUB-2017-011, CERN, 2017, URL: https://cds.cern.ch/record/2270366 (cit. on p. 50).

[103] ATLAS Collaboration, *Topological b-hadron decay reconstruction and identification of b-jets with the JetFitter package in the ATLAS experiment at the LHC*, tech. rep. ATL-PHYS-PUB-2018-025, CERN, 2018, URL: https://cds.cern.ch/record/2645405 (cit. on p. 50).

[104] A. Hoecker et al., *TMVA - Toolkit for Multivariate Data Analysis*, 2007, arXiv: physics/0703039 [physics.data-an] (cit. on pp. 52, 149, 150).

[105] ATLAS Collaboration, *Optimisation and performance studies of the ATLAS b-tagging algorithms for the 2017-18 LHC run*, tech. rep. ATL-PHYS-PUB-2017-013, CERN, 2017, URL: https://cds.cern.ch/record/2273281 (cit. on p. 52).

[106] ATLAS Collaboration, *Expected performance of the 2019 ATLAS b-taggers*, Accessed: 2020-11-12, URL: https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2019-005/ (cit. on pp. 52, 128).

[107] ATLAS Collaboration, *Calibration of the ATLAS b-tagging algorithm in $t\bar{t}$ semi-leptonic events*, tech. rep. ATLAS-CONF-2018-045, CERN, 2018, URL: http://cds.cern.ch/record/2638455 (cit. on p. 52).

[108] ATLAS Collaboration, *Measurement of b-tagging Efficiency of c-jets in $t\bar{t}$ Events Using a Likelihood Approach with the ATLAS Detector*, tech. rep. ATLAS-CONF-2018-001, CERN, 2018, URL: http://cds.cern.ch/record/2306649 (cit. on p. 52).

[109] ATLAS Collaboration, *Calibration of light-flavour b-jet mistagging rates using ATLAS proton-proton collision data at $\sqrt{s} = 13$ TeV*, tech. rep. ATLAS-CONF-2018-006, CERN, 2018, URL: https://cds.cern.ch/record/2314418 (cit. on p. 52).

[110] ATLAS Collaboration, *Monte Carlo to Monte Carlo scale factors for flavour tagging efficiency calibration*, tech. rep. ATL-PHYS-PUB-2020-009, CERN, 2020, URL: https://cds.cern.ch/record/2718610 (cit. on p. 52).

[111] ATLAS Collaboration, *Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using pp collisions at $\sqrt{s} = 13$ TeV*, tech. rep. ATLAS-CONF-2017-029, CERN, 2017, URL: http://cds.cern.ch/record/2261772 (cit. on p. 53).

[112] ATLAS Collaboration, *Measurement of the Higgs boson decaying to b-quarks produced in association with a top-quark pair in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector*, tech. rep. ATLAS-CONF-2020-058, CERN, 2020, URL: http://cds.cern.ch/record/2743685 (cit. on p. 56).

[113] ATLAS Collaboration, *Performance of electron and photon triggers in ATLAS during LHC Run 2*, Eur. Phys. J. **C80** (2020) 47, arXiv: 1909.00761 [hep-ex] (cit. on pp. 56, 57).

[114] ATLAS Collaboration, *Performance of the ATLAS muon triggers in Run 2*, 2020, arXiv: 2004.13447 [hep-ex] (cit. on pp. 56, 57).

[115] T. Sjöstrand et al., *An Introduction to PYTHIA 8.2*, Comput. Phys. Commun. **191** (2015) 159, arXiv: 1410.3012 [hep-ph] (cit. on pp. 56, 59).

[116] G. Aad, B. Abbott, J. Abdallah, O. Abdinov, and B. Abeloos, *Charged-particle distributions in s=13 TeV pp interactions measured with the ATLAS detector at the LHC*, Physics Letters B **758** (2016) 67, ISSN: 0370-2693, URL: http://www.sciencedirect.com/science/article/pii/S037026931630123X (cit. on p. 56).

[117] ATLAS Collaboration, *The Pythia 8 A3 tune description of ATLAS minimum bias and inelastic measurements incorporating the Donnachie-Landshoff diffractive model*, tech. rep. ATL-PHYS-PUB-2016-017, CERN, 2016, URL: https://cds.cern.ch/record/2206965 (cit. on p. 56).

[118] ATLAS Collaboration, *ATLAS Pythia 8 tunes to 7 TeV datas*, tech. rep. ATL-PHYS-PUB-2014-021, CERN, 2014, URL: https://cds.cern.ch/record/1966419 (cit. on pp. 56, 59, 61).

[119] M. Bahr et al., *Herwig++ Physics and Manual*, Eur. Phys. J. C **58** (2008) 639, arXiv: 0803.0883 [hep-ph] (cit. on pp. 56, 60).

[120] D. J. Lange, *The EvtGen particle decay simulation package*, Nucl. Instrum. Meth. A **462** (2001) 152 (cit. on p. 57).

[121] ATLAS Collaboration, *Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector*, Eur. Phys. J. C **75** (2015) 349, arXiv: 1503.05066 [hep-ex] (cit. on p. 57).

[122] ATLAS Collaboration, *Measurements of inclusive and differential fiducial cross-sections of $t\bar{t}$ production with additional heavy-flavour jets in proton-proton collisions at $\sqrt{s}$ = 13 TeV with the ATLAS detector*, JHEP **04** (2019) 046, arXiv: 1811.12113 [hep-ex] (cit. on pp. 57, 89, 115, 128).

[123] T. Ježo, J. M. Lindert, N. Moretti, and S. Pozzorini, *New NLOPS predictions for $t\bar{t}+b$ -jet production at the LHC*, Eur. Phys. J. C **78** (2018) 502, arXiv: 1802.00426 [hep-ph] (cit. on p. 59).

[124] F. Cascioli, P. Maierhofer, and S. Pozzorini, *Scattering Amplitudes with Open Loops*, Phys. Rev. Lett. **108** (2012) 111601, arXiv: 1111.5206 [hep-ph] (cit. on p. 59).

[125] A. Denner, S. Dittmaier, and L. Hofer, *Collier: a fortran-based Complex One-Loop LIbrary in Extended Regularizations*, Comput. Phys. Commun. **212** (2017) 220, arXiv: 1604.06792 [hep-ph] (cit. on p. 59).

[126] R. D. Ball et al., *Parton distributions for the LHC run II*, Journal of High Energy Physics **2015** (2015), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP04(2015)040 (cit. on p. 59).

[127] J. Alwall et al., *The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations*, JHEP **07** (2014) 079, arXiv: 1405.0301 [hep-ph] (cit. on p. 61).

[128] P. Nason, *A New method for combining NLO QCD with shower Monte Carlo algorithms*, JHEP **11** (2004) 040, arXiv: hep-ph/0409146 (cit. on p. 61).

[129] H. B. Hartanto, B. Jager, L. Reina, and D. Wackeroth, *Higgs boson production in association with top quarks in the POWHEG BOX*, Phys. Rev. D **91** (2015) 094003, arXiv: 1501.04498 [hep-ph] (cit. on p. 61).

[130] NNPDF Collaboration, R.D. Ball et al., *Parton distributions for the LHC Run II*, JHEP **04** (2015) 040, arXiv: 1410.8849 [hep-ph] (cit. on p. 61).

[131] ATLAS Collaboration, *Selection of jets produced in 13TeV proton-proton collisions with the ATLAS detector*, tech. rep. ATLAS-CONF-2015-029, CERN, 2015, URL: https://cds.cern.ch/record/2037702 (cit. on p. 63).

[132] T. P. Calvet, *Search for the production of a Higgs boson in association with top quarks and decaying into a b-quark pair and b-jet identification with the ATLAS experiment at LHC*, (2017), Presented 08 Nov 2017, URL: https://cds.cern.ch/record/2296985 (cit. on p. 66).

[133] I. Antcheva et al., *ROOT — A C++ framework for petabyte data storage, statistical analysis and visualization*, Computer Physics Communications **180** (2009) 2499, 40 YEARS OF CPC: A celebratory issue focused on quality software for high performance, grid and novel computing architectures, ISSN: 0010-4655, URL: http://www.sciencedirect.com/science/article/pii/S0010465509002550 (cit. on pp. 67, 81, 149).

[134] *Data Analysis Techniques for High Energy Particle Physics*, CERN, Proceedings of the 1974 CERN School of Computing: Godøysund, Norway 11 - 24 Aug 1974. 3rd CERN School of Computing, 1974, URL: https://cds.cern.ch/record/186223 (cit. on p. 68).

[135] Y. Zhang, W.-G. Ma, R.-Y. Zhang, C. Chen, and L. Guo, *QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider*, Phys. Lett. B **738** (2014) 1, arXiv: 1407.1110 [hep-ph] (cit. on p. 73).

[136] LHCTopWG, *ATLAS-CMS recommended predictions for top-quark-pair cross sections*, URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO (cit. on p. 73).

[137] ATLAS Collaboration, *Measurement of the Inelastic Proton–Proton Cross Section at $\sqrt{s} = 13\,TeV$ with the ATLAS Detector at the LHC*, Phys. Rev. Lett. **117** (2016) 182002, arXiv: 1606.02625 [hep-ex] (cit. on p. 73).

[138] G. Cowan, *Statistical data analysis*, Oxford University Press, USA, 1998 (cit. on pp. 77, 78, 81, 89).

[139] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, The European Physical Journal C **71** (2011), ISSN: 1434-6052, URL: http://dx.doi.org/10.1140/epjc/s10052-011-1554-0 (cit. on pp. 77, 79).

[140] R. Cousins, "Generalization of Chisquare Goodness-ofFit Test for Binned Data Using Saturated Models , with Application to Histograms", 2013, URL: http://www.physics.ucla.edu/~cousins/stats/cousins_saturated.pdf (cit. on pp. 81, 103).

[141] K. Cranmer, G. Lewis, L. Moneta, A. Shibata, and W. Verkerke, *HistFactory: A tool for creating statistical models for use with RooFit and RooStats*, tech. rep. CERN-OPEN-2012-016, New York U., 2012, URL: https://cds.cern.ch/record/1456844 (cit. on p. 81).

[142] W. Verkerke and D. Kirkby, *The RooFit toolkit for data modeling*, 2003, arXiv: physics/0306116 [physics.data-an] (cit. on p. 81).

[143] L. Moneta et al., *The RooStats Project*, 2010, arXiv: 1009.1003 [physics.data-an] (cit. on p. 81).

[144] F. James and M. Roos, *Minuit - a system for function minimization and analysis of the parameter errors and correlations*, Computer Physics Communications **10** (1975) 343, ISSN: 0010-4655, URL: http://www.sciencedirect.com/science/article/pii/0010465575900399 (cit. on p. 81).

[145] B. Efron, *Bootstrap Methods: Another Look at the Jackknife*, Ann. Statist. **7** (1979) 1, URL: https://doi.org/10.1214/aos/1176344552 (cit. on p. 94).

[146] ATLAS Collaboration, *In situ calibration of large-radius jet energy and mass in 13 TeV proton–proton collisions with the ATLAS detector*, Eur. Phys. J. C **79** (2019) 135, arXiv: 1807.09477 [hep-ex] (cit. on p. 106).

[147] N. Scharmberg, *Measurement of the production cross section of a Higgs boson in combination with two top quarks at the ATLAS experiment*, (2020), Thesis in preparation (cit. on pp. 107, 109, 110).

[148] R. Frederix and S. Frixione, *Merging meets matching in MC@NLO*, Journal of High Energy Physics **2012** (2012), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP12(2012)061 (cit. on p. 128).

[149] S. Höche, F. Krauss, S. Schumann, and F. Siegert, *QCD matrix elements and truncated showers*, Journal of High Energy Physics **2009** (2009) 053, ISSN: 1029-8479, URL: http://dx.doi.org/10.1088/1126-6708/2009/05/053 (cit. on p. 128).

[150] S. Höche, F. Krauss, M. Schönherr, and F. Siegert, *QCD matrix elements + parton showers. The NLO case*, Journal of High Energy Physics **2013** (2013), ISSN: 1029-8479, URL: http://dx.doi.org/10.1007/JHEP04(2013)027 (cit. on p. 128).

[151] S. Höche, J. Krause, and F. Siegert, *Multijet merging in a variable flavor number scheme*, Physical Review D **100** (2019), ISSN: 2470-0029, URL: http://dx.doi.org/10.1103/PhysRevD.100.014011 (cit. on p. 128).

[152] *1st Terascale School of Machine Learning*, Accessed: 2020-08-16, URL: https://indico.desy.de/indico/event/21278/overview (cit. on p. 149).

[153] N. A. Asbah, *Search for the Production of a Standard Model Higgs Boson in Association with Top-Quarks and Decaying into a Pair of Bottom-Quarks with 13 TeV ATLAS Data*, (2018), Presented 23 May 2018, URL: https://cds.cern.ch/record/2320703 (cit. on pp. 149, 150).

[154] J. A. Raine, *Evidence for the production of a Higgs boson in association with two top quarks with the ATLAS detector: A search in the $H \to b\bar{b}$ channel and in combination with other Higgs boson decays at $\sqrt{s} = 13$ TeV*, (2018), Presented 26 Mar 2018, URL: https://cds.cern.ch/record/2316953 (cit. on p. 150).

List of Figures

List of Tables

List of acronyms

E_{T}^{miss} missing E_{T}. 37, 73

BDT Boosted Decission Tree. xiv, 52, 53, 65, 108, 149, 150

BSM Beyond the Standard Model. 26

CSC Cathode Strip Chambers. 34

DNN Deep Neural Network. 52

EM electromagnetic. 3, 5, 6, 7, 9, 27, 31, 32, 33, 39, 42, 43, 44, 48, 49, 140, 141

EW electroweak. xi, 3, 4, 5, 6, 7, 8

FSR Final State Radiation. 17, 59, 60, 61, 71, 81, 94, 99, 96, 99, 101, 140, 155, 159

HEP High Energy Physics. 28

HLT high-level trigger. 37, 38

IBL Insertable B-layer. 30, 37

ID Inner Detector. 27, 28, 31, 32, 34, 39, 40, 41, 42, 43, 44, 45, 46, 50, 53, 140, 141

IP impact parameter. 28, 44

ISR Initial State Radiation. 17, 19, 59, 60, 61, 71, 81, 83, 85, 87, 94, 99, 96, 99, 101, 111, 114, 115, 116, 115, 119, 124, 122, 126, 140, 143, 144, 154, 158

JER jet energy resolution. 49, 73

JES jet energy scale. 49, 73

JVT jet vertex tagger. 49, 63, 73

L1 Level-1. 37, 38

LAr liquid argon. 32, 34

LH likelihood. 43, 44

LHC Large Hadron Collider. 2, 3, 11, 23, 25, 26, 27, 28, 37, 38, 55

LO leading order. 15, 17, 20, 22, 140

MC Monte Carlo. 14, 21, 39, 45, 48, 49, 52, 54, 56, 57, 74, 119, 122, 127, 162

MDT Monitored Drift Tube. 34, 36

MIP minimum ionizing particle. 45, 46

MS Muon Spectometer. 27, 39, 45, 46

MVA multi-variate algorithm. vii, 52, 65, 108, 127, 150, 151

NLO next-to-leading order. 15, 17, 19, 20, 21, 22, 58, 61, 72, 92, 126, 128, 140, 141

NLO match NLO matching. 59, 61, 68, 72, 81, 83, 85, 89, 92, 95, 96, 99, 101, 103, 111, 117, 118, 119, 122, 124, 122, 126, 141, 157, 161

PDF parton distribution function. 14, 15, 19, 59, 60, 61

PS parton shower. 56, 61, 68, 71, 83, 96, 99, 111, 128, 141

PS&had Parton shower & hadronization. 59, 60, 61, 71, 72, 81, 83, 95, 96, 99, 101, 117, 156, 160

PV primary vertex. 42, 48, 49, 50, 63

QCD Quantum Chromodynamics. 3, 4, 8, 9, 15, 17, 20, 21, 46, 61, 140

QED Quantum Electrodynamics. xi, 3, 5, 6

RPC Resistive Plate Chambers. 34, 36

SCT Semiconductor Tracker. 28, 30, 31, 40, 41, 40, 41, 141

SM Standard Model. vii, xi, 1, 3, 4, 5, 8, 9, 10, 11, 13, 14, 26, 31, 54, 76, 106, 108, 118, 124, 127

TGC Thin Gap Chambers. 34, 36

TRT Transition Radiation Tracker. 28, 31, 34, 41, 43

UE underlying event. 21

WP working point. 44, 45, 46, 52

APPENDIX A

Multivariate algorithms and their application

This appendix offers a short overview of the Boosted Decission Tree (BDT) and its implementation throughout the analysis presented in this thesis. Most of the information presented in this appendix is based on the lectures presented in references [152, 153].

A.1 Boosted decision trees

Decision tress (DT) offer a way for classification of events. They are based on binary trees, where in each node events are separated into two groups corresponding to each of the classification categories. This division is done by selection on one of the input variables. The best selection criteria are determined by a separation function and in each node the best separating variable is chosen.

The events are further divided until a stopping criterion is reached - it can be a minimal number of events to avoid statistical fluctuations, number of divisions or simply reaching desired separation (e.g. when more than 90% of events in given node fall into one of the categories).

In a simple example, one can imagine classification between the signal and background. In the first node, the variable which best separates between the two is chosen, separating the input events into two groups, one more signal-like and one more background-like. Then the process is repeated on each of the groups, ideally separating events into groups of mostly signal or mostly background. In each node a new variable is chosen in order to get the best separation. An example of a DT can be found in figure A.1.

Boosted decision trees (BDT) are decision trees with "boosting", where a large number of trees is created iteratively, the next trees created by putting larger weight on events wrongly classified. All the trees created are then combined into a single and better classifier.

Implementation of the BDTs used in the $t\bar{t}H(b\bar{b})$ analysis is based on the TMVA package[104] implemented in the ROOT data analysis framework[133].

A.2 Reconstruction BDT

The reconstruction BDT used in the $t\bar{t}H(b\bar{b})$ analysis aims to find the best match between the partons of the $t\bar{t}H(b\bar{b})$ process and the reconstructed objects. First, candidates for each parton are built based on their expected decay products. This means for example

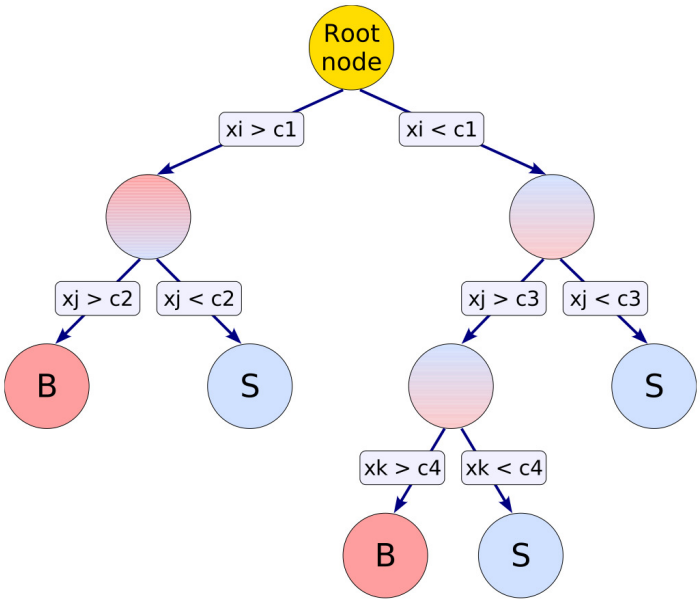


Figure A.1: An example of a boosted decision tree[104]. In each node the events are divided using some feature x based on whether they are more signal-like or background-like, until some stopping condition is reached.

combination of two b-jets as a Higgs candidate, or two jets as a hadronic W candidate. The BDT is then trained to distinguish between the wrong and correct combinations.

As an input, particle properties of the reconstructed candidates are used, e.g. Higgs p_{T} or mass of a reconstructed top quark. In case of the resolved channels, two categories are trained, one assuming the presence of a Higgs boson in the event and one not. This is necessary since the $t\bar{t}b\bar{b}$ final state does not contain any and the assumption that there is one could lead to larger probability of wrong parton assignment.

The implementation of this BDT did not change since the previous iteration of the analysis[8], where a more detailed description can be found in reference [153] for the single lepton regions and in reference [154] for the dilepton regions.

A.3 Classification BDT

The goal of the classification BDT is to separate the signal ($t\bar{t}H$) from the background, which is dominated by the $t\bar{t}b\bar{b}$ process but contains also small contributions of the $t\bar{t}+\geq 1c$ and $t\bar{t}$+light final states. It is trained separately in different analysis regions and channels.

For the single lepton channel a detailed description can be found in reference [153]. It uses several kinematic variables and outputs of other MVAs, e.g. the reconstruction BDT. The results shown in this thesis are performed using the classification BDT as used in the previous analysis[8]. This means it is trained on the $t\bar{t}$+jets sample as described in section 6.4.3 and not on the nominal $t\bar{t}b\bar{b}$ sample. A retraining was attempted but the new BDT showed a mis-modeling with respect to the data which was never resolved.

In the dilepton channel the classification BDT was retrained using the $t\bar{t}b\bar{b}$ sample, improving its performance. The implementation, despite the different background sample, is the same as in the previous iteration fo the analysis[8], which is described in more detail in reference [154].

APPENDIX B

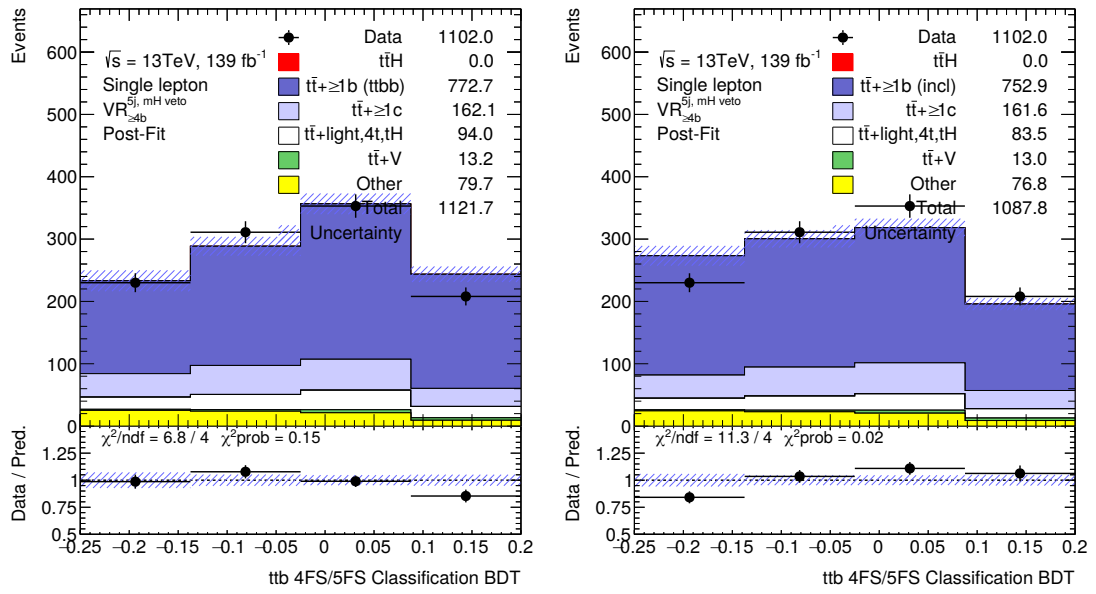
Choice of the nominal $t\bar{t}b\bar{b}$ model

The choice of the nominal sample is an important part of any analysis. In the case of the $t\bar{t}H(b\bar{b})$ analysis presented in this thesis, two samples were considered: the $t\bar{t}b\bar{b}$ and the $t\bar{t}$+jets sample. Both were generated using PowhegBox +Pythia8, with more details given in section 6.4.3.

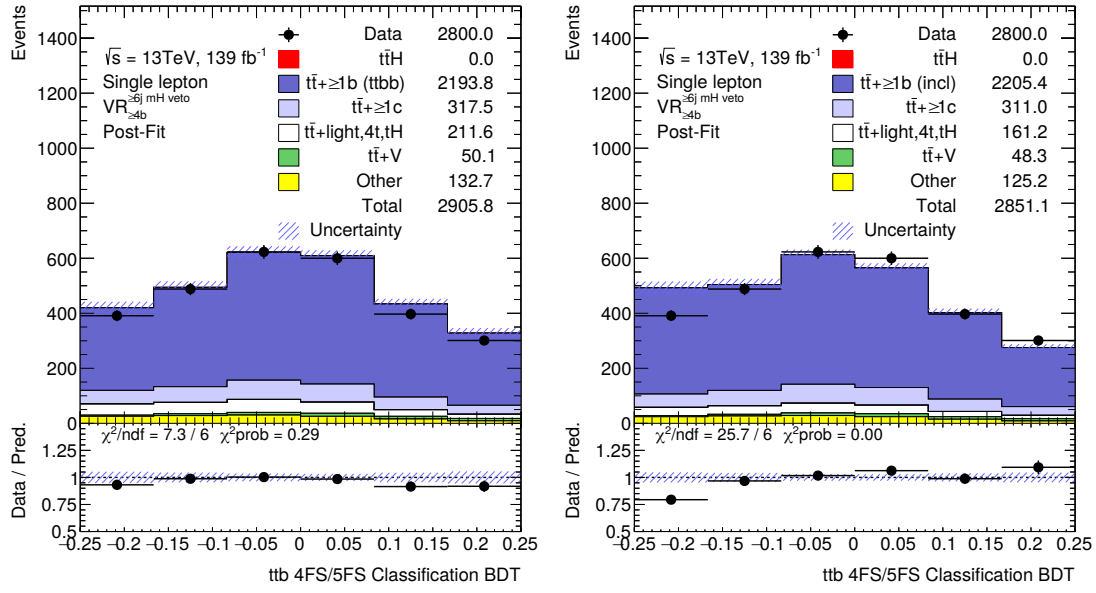
The choice of the nominal sample was based on a study made by Ryunosuke Iguchi. First, a special MVA was trained to discriminate between the two samples.

The background-only fit, as described in section 7.5, was then performed with both samples as the nominal $t\bar{t}+\geq 1b$ background. The post-fit values of the nuisance parameter were applied in two validation regions: $\mathbf{VR}_{\geq 4b}^{5j,m_H \text{ veto}}$ which requires exactly 5 jets and $\mathbf{VR}_{\geq 4b}^{6j,m_H \text{ veto}}$ which requires at least 6 jets. Both selections require 4 b-jets at the 70% working point and events with reconstructed Higgs mass between 85 and 145 GeV are removed to reduce contribution of the signal. The Higgs mass was determined using the reconstruction BDT described in appendix A.

Both regions for both candidates for the nominal sample are displayed in figure B.1, showing that the $t\bar{t}b\bar{b}$ sample provides a better agreement with the data.



(a)



(b)

Figure B.1: Distributions of the two validation region used to determined the nominal $t\bar{t}+\geq 1b$ sample, (a) the $\mathrm{VR}^{5j,m_H\ \mathrm{veto}}_{\geq 4b}$ region and (b) the $\mathrm{VR}^{6j,m_H\ \mathrm{veto}}_{\geq 4b}$ region. On the left is the $t\bar{t}b\bar{b}$ sample while on the right the $t\bar{t}$+jets sample, all shown as post-fit plots of a background fit with the corresponding sample used as the $t\bar{t}+\geq 1b$ component. Courtesy of Ryunosuke Iguchi.

APPENDIX C

Distributions of modeling systematic variations

In this appendix, shapes of the modeling systematic uncertainties of the signal $t\bar{t}H$ and the dominant background $t\bar{t}b\bar{b}$ (see sections 6.4.4 and 6.4.3 respectively) are presented in the four analysis regions described in section 6.5.2. Sources of the uncertainties were discussed in section 6.4.4 for the $t\bar{t}H$ process and in section 6.4.3 for the $t\bar{t}b\bar{b}$ sample. The variations are modified by the PARABOLIC smoothing and symmetrization procedure described previously in section 6.7.

C.1 Variations of the $t\bar{t}b\bar{b}$ modeling

$t\bar{t}+\geq 1b$ ISR systematic

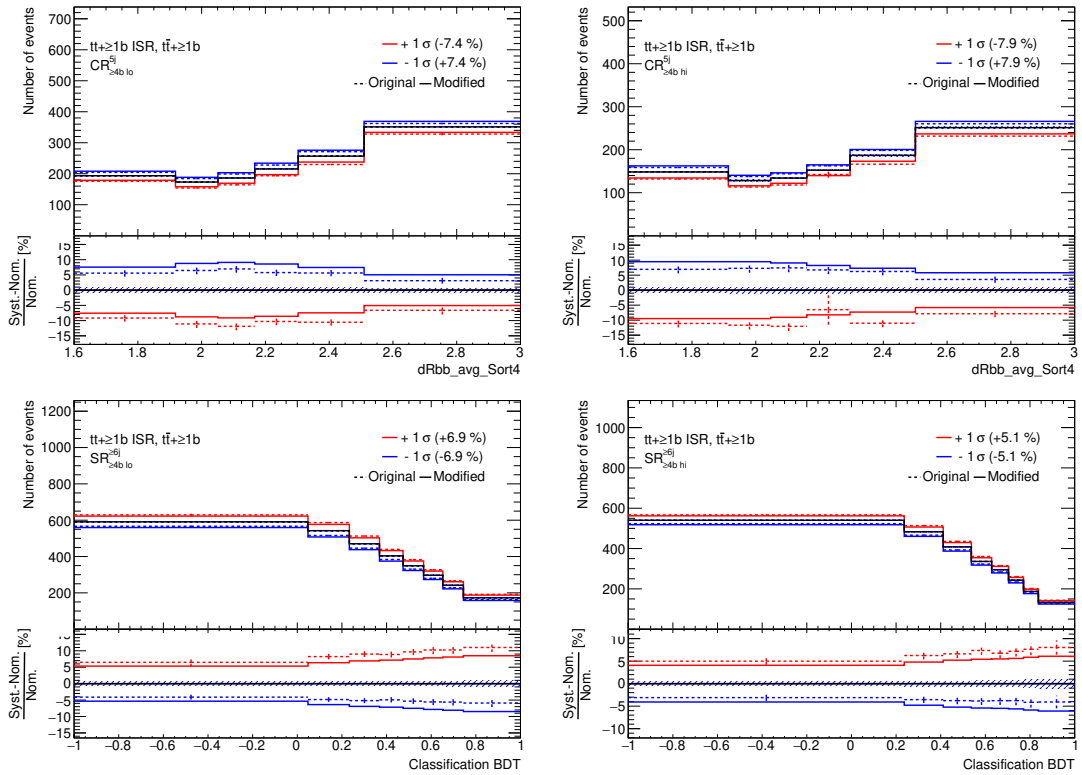


Figure C.1: Distributions of the $t\bar{t}+\geq 1b$ ISR systematic in all analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and the smoothing.

$t\bar{t}+\geq 1b$ FSR systematic

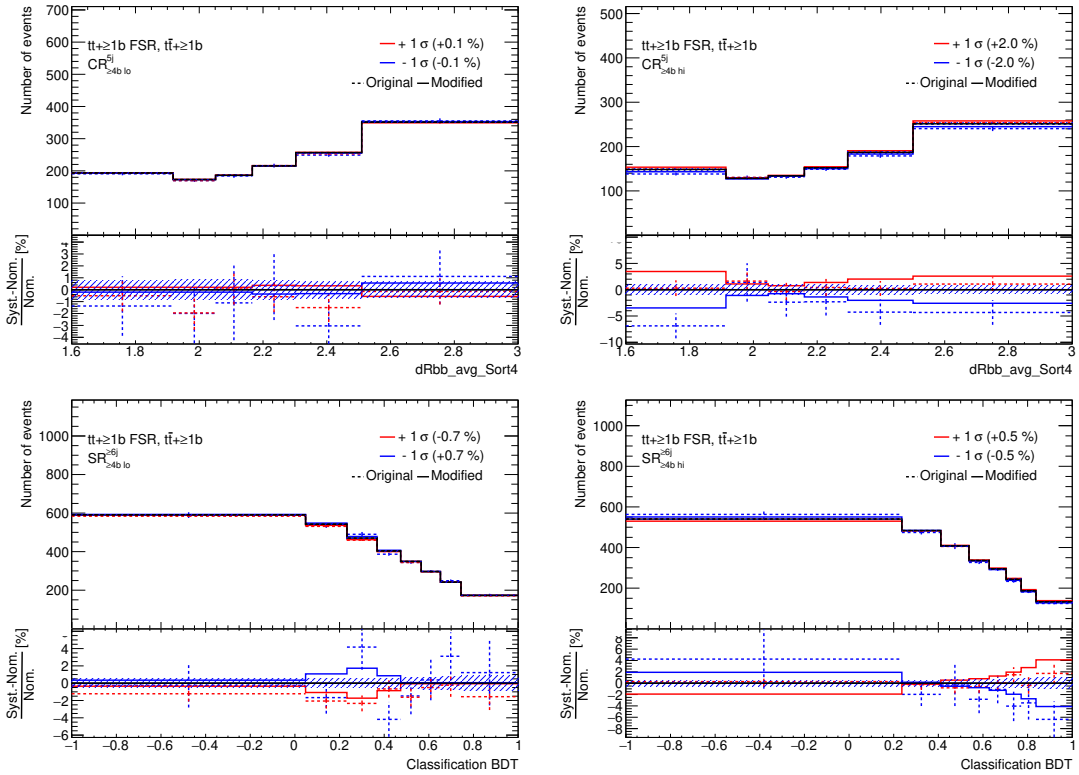


Figure C.2: Distributions of the $t\bar{t}+\geq 1b$ FSR systematic in all the analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and smoothing.

$t\bar{t}+\geq 1b$ PS&had systematic

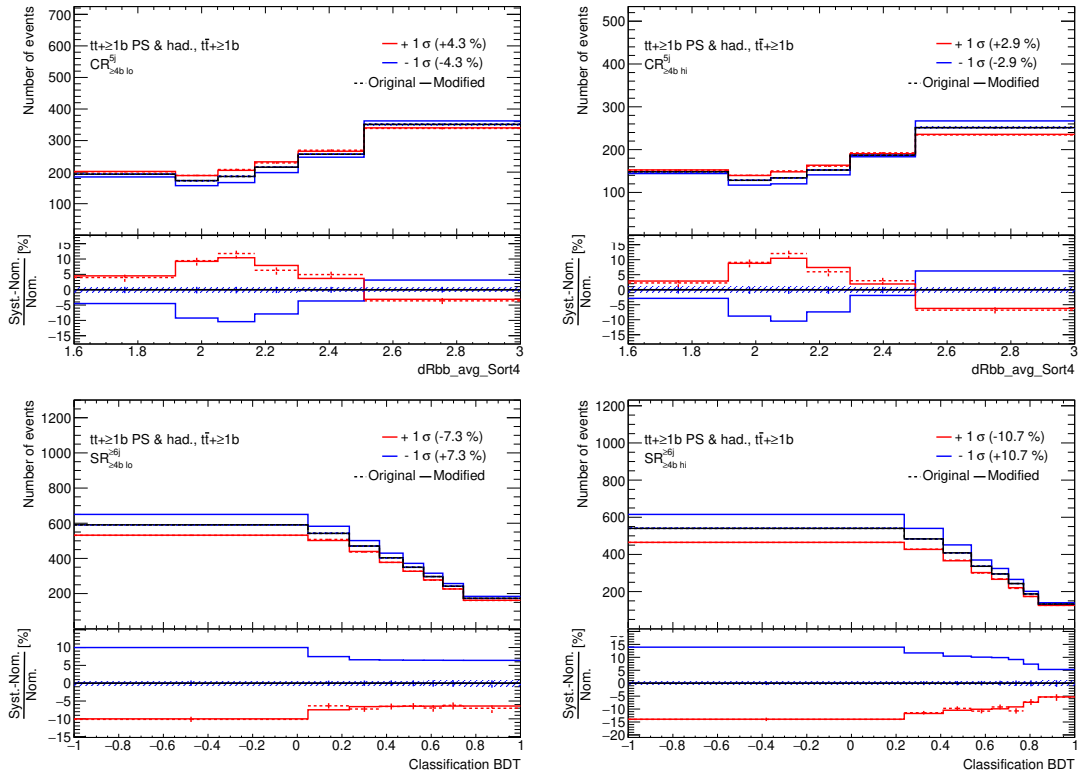


Figure C.3: Distributions of the $t\bar{t}+\geq 1b$ Parton shower & hadronization systematic in all the analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and smoothing.

$t\bar{t}+\geq 1b$ NLO match systematic

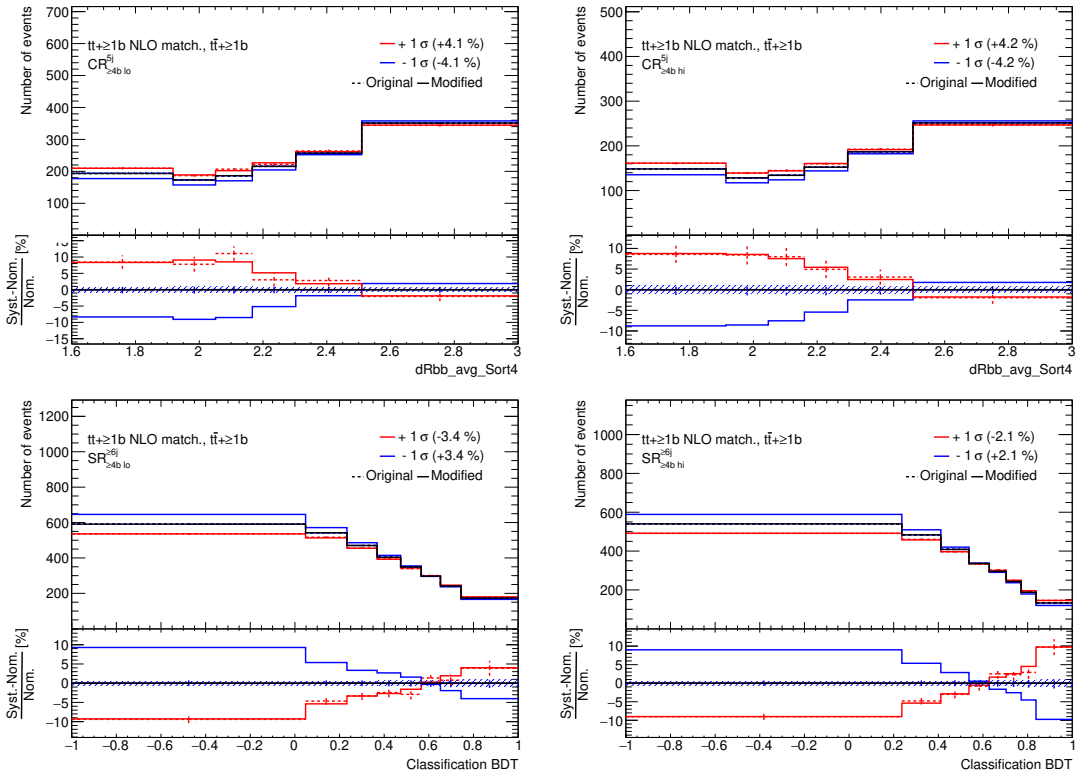


Figure C.4: Distributions of the $t\bar{t}+\geq 1b$ NLO matching systematic in all the analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and smoothing.

C.2 Variations of the $t\bar{t}H$ modeling

$t\bar{t}H$ ISR systematic

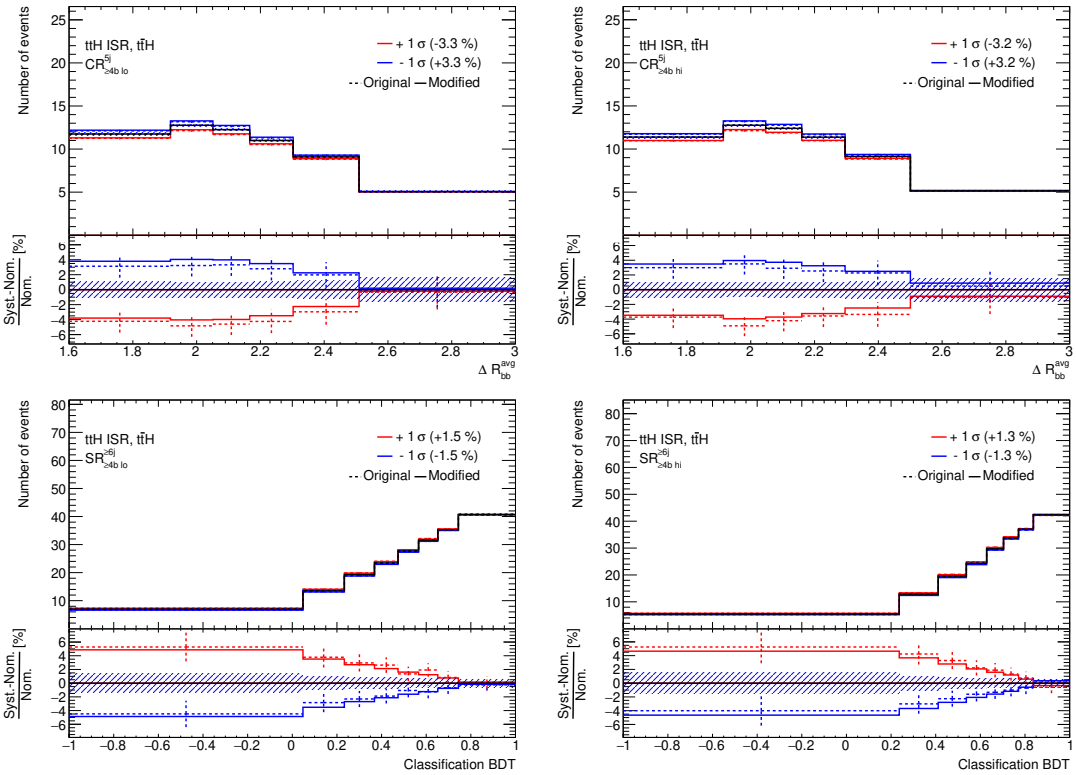


Figure C.5: Distributions of the $t\bar{t}H$ ISR systematic in all analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and the smoothing.

$t\bar{t}H$ FSR systematic

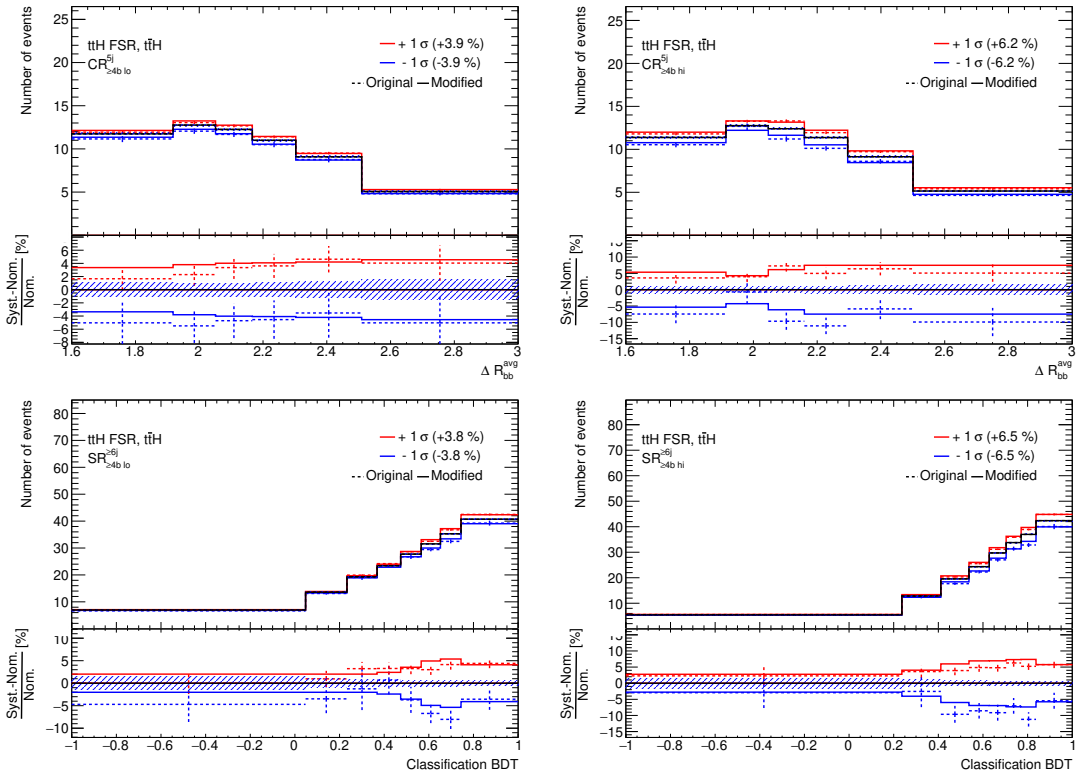


Figure C.6: Distributions of the $t\bar{t}H$ FSR systematic in all the analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and smoothing.

$t\bar{t}H$ PS&had systematic

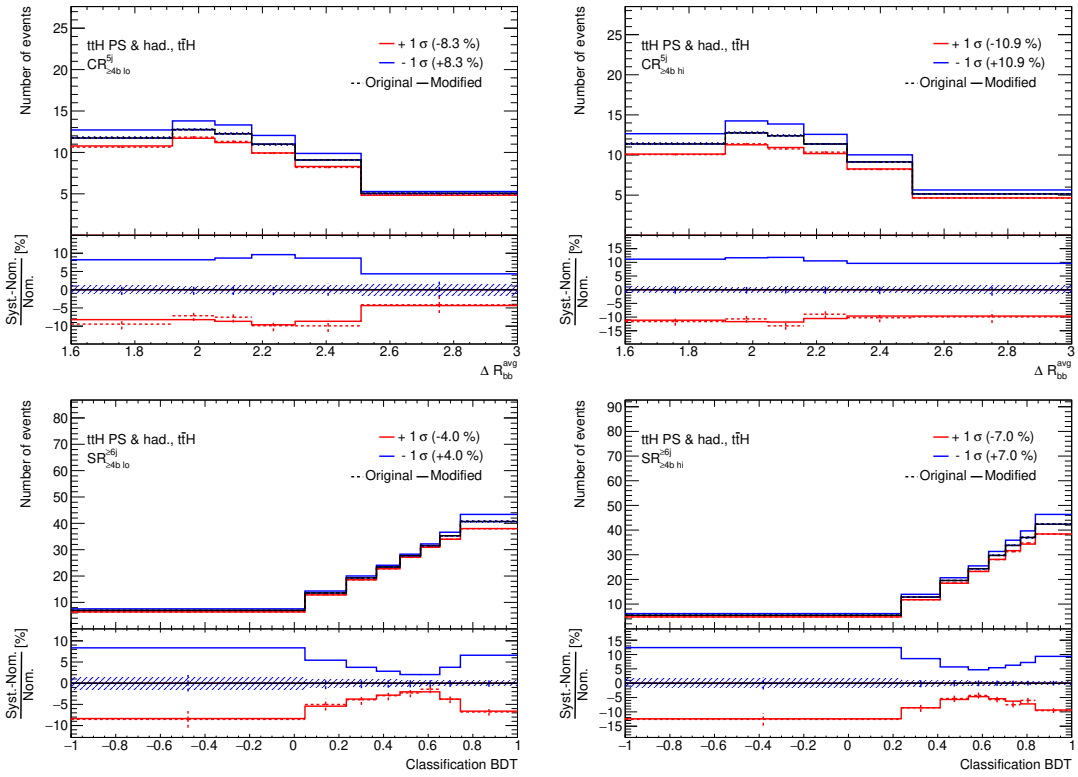


Figure C.7: Distributions of the $t\bar{t}H$ Parton shower & hadronization systematic in all the analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and smoothing.

$t\bar{t}H$ NLO match systematic

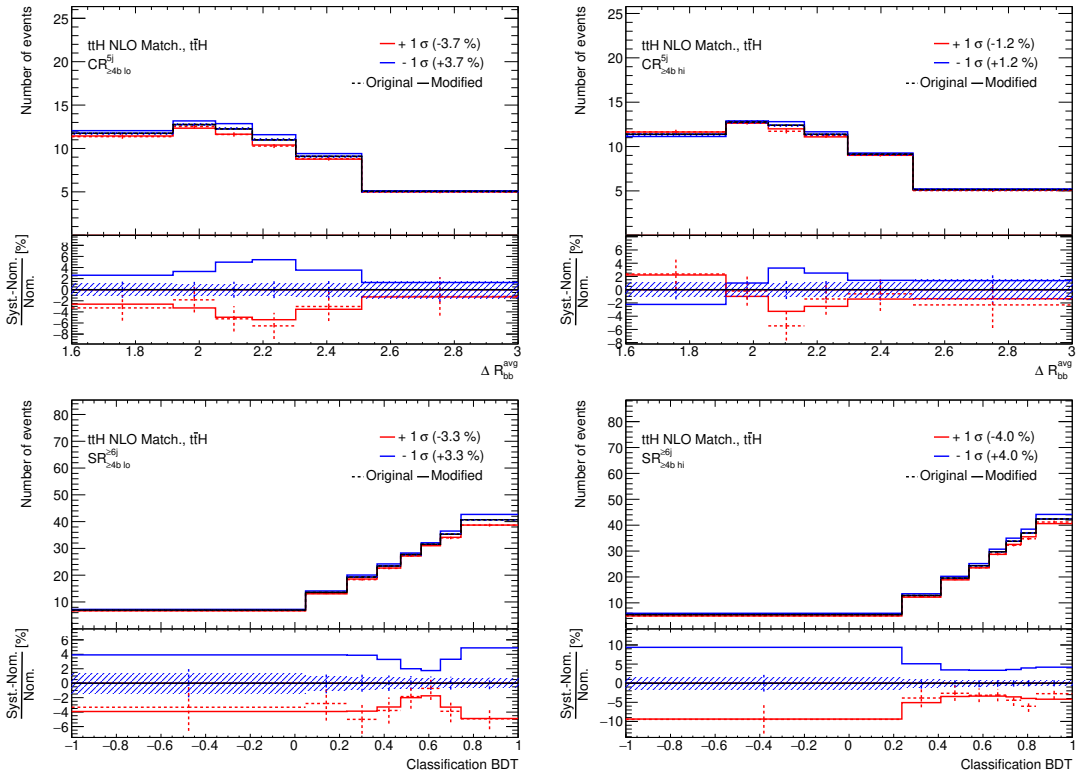


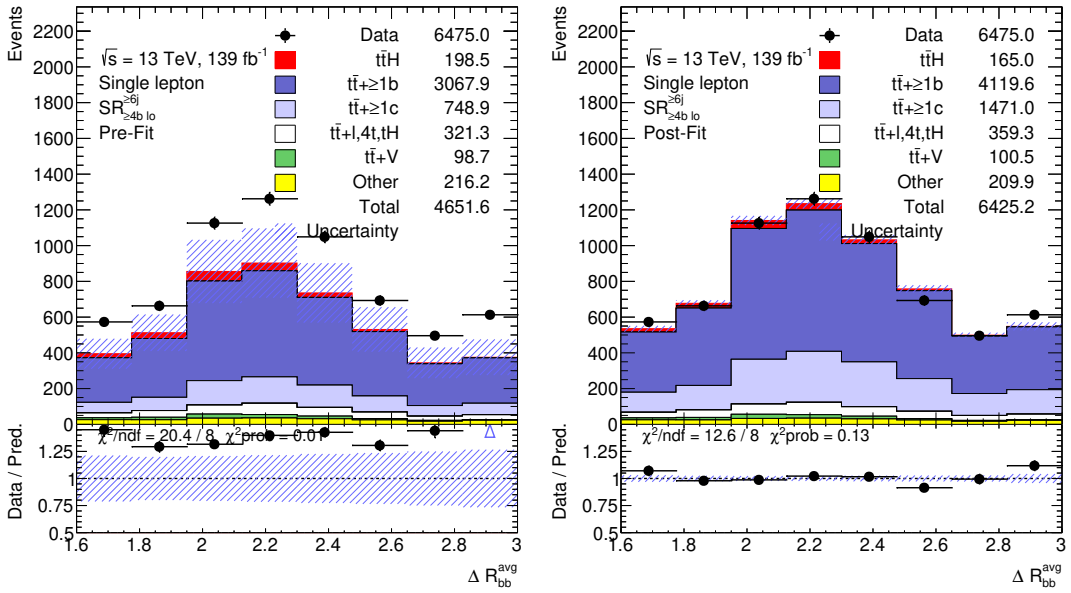
Figure C.8: Distributions of the $t\bar{t}H$ NLO matching systematic in all the analysis regions, at the top $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{CR}^{5j}_{\geq 4b\ \mathrm{hi}}$ (right) and at the bottom $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ (left) and $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ (right). Original refers to the raw input distribution, modified is the distribution after the symmetrization and smoothing.

APPENDIX D

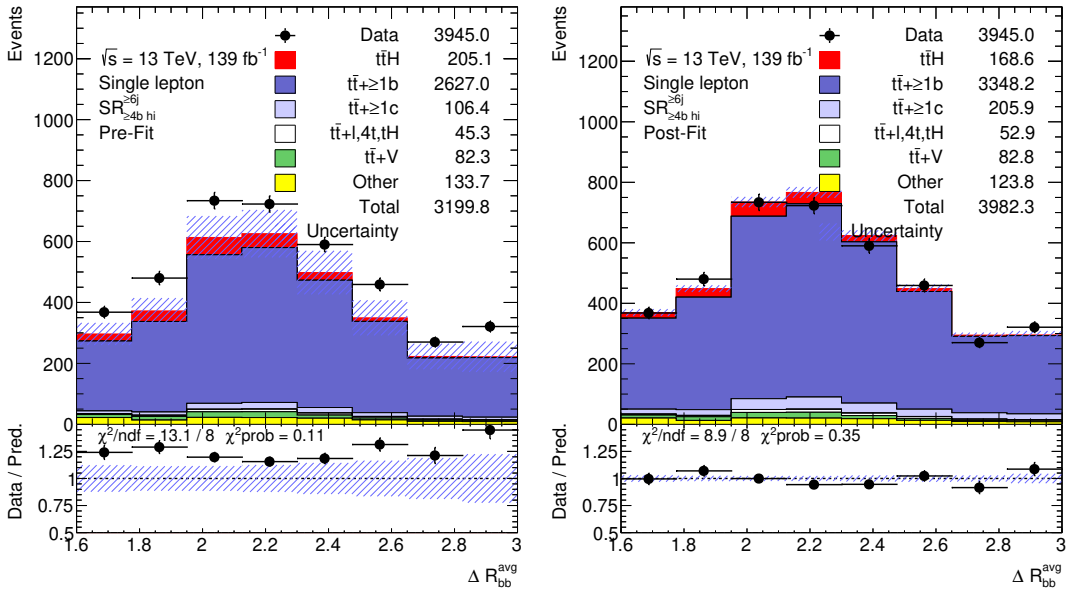
Modeling of other variables

This appendix presents comparisons of the data and the MC in the four analysis regions of the single lepton channel. It is done for a series of variables not used in the fit. Two of them are angular variables $\Delta R_{bb}^{\text{avg}}$ (which is used in the fit only in the control regions) and $\Delta\eta_{\text{jj}}^{\text{max}}$ which gives the distance in η between the two most distant jets. In addition, a Higgs mass $m_H^{\text{reco BDT}}$, determined using the reconstruction BDT (see section A), is shown. None of these variable shows a significant post-fit mis-modeling, with a possible slope in the $m_H^{\text{reco BDT}}$ variable for both signal regions (shown in figure D.5). Figure compare the pre-fit distribution and post-fit distribution for the combined measurement.

Distributions of the $\Delta R_{bb}^{\text{avg}}$ variable



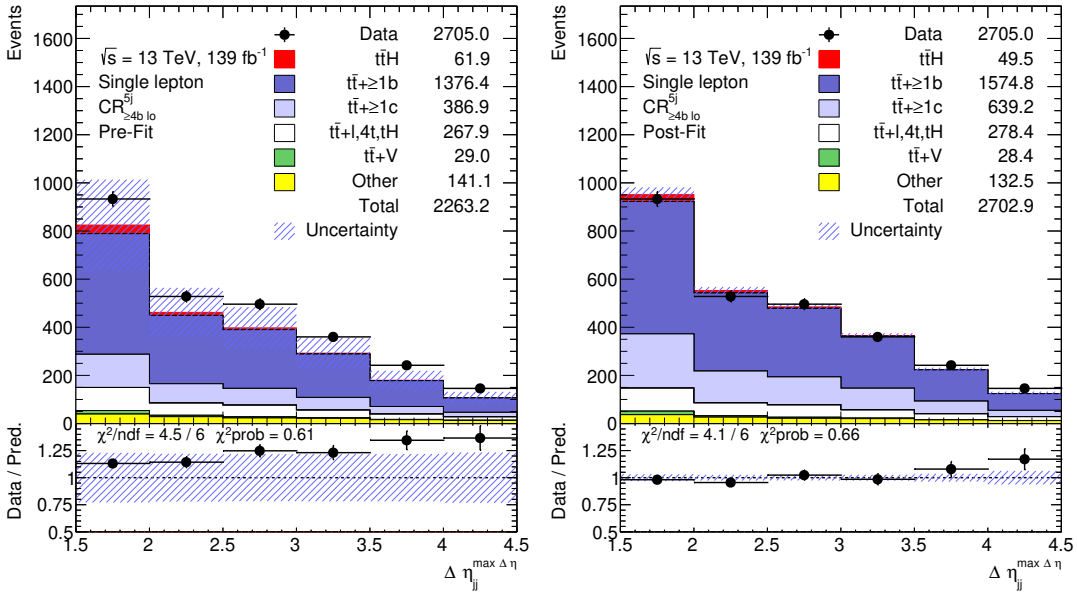
(a)



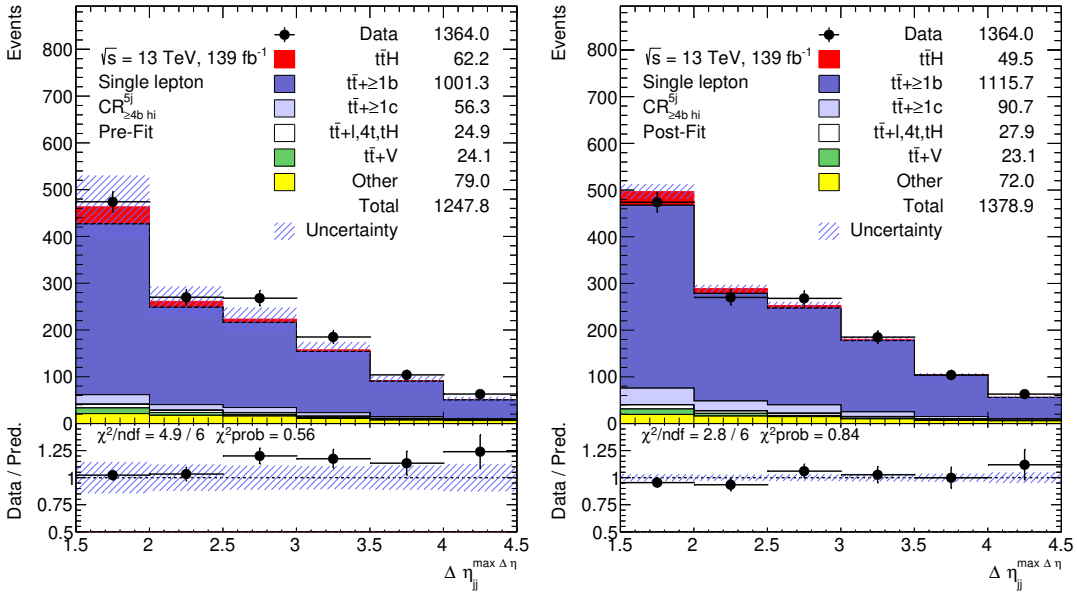
(b)

Figure D.1: Distributions of $\Delta R_{bb}^{\text{avg}}$ pre-fit (left) and post-fit (right) in the (a) $\text{SR}^{\geq 6j}_{\geq 4b\ \text{lo}}$ and the (b) $\text{SR}^{\geq 6j}_{\geq 4b\ \text{hi}}$ regions.

Distributions of the $\Delta\eta_{jj}^{max}$ variable

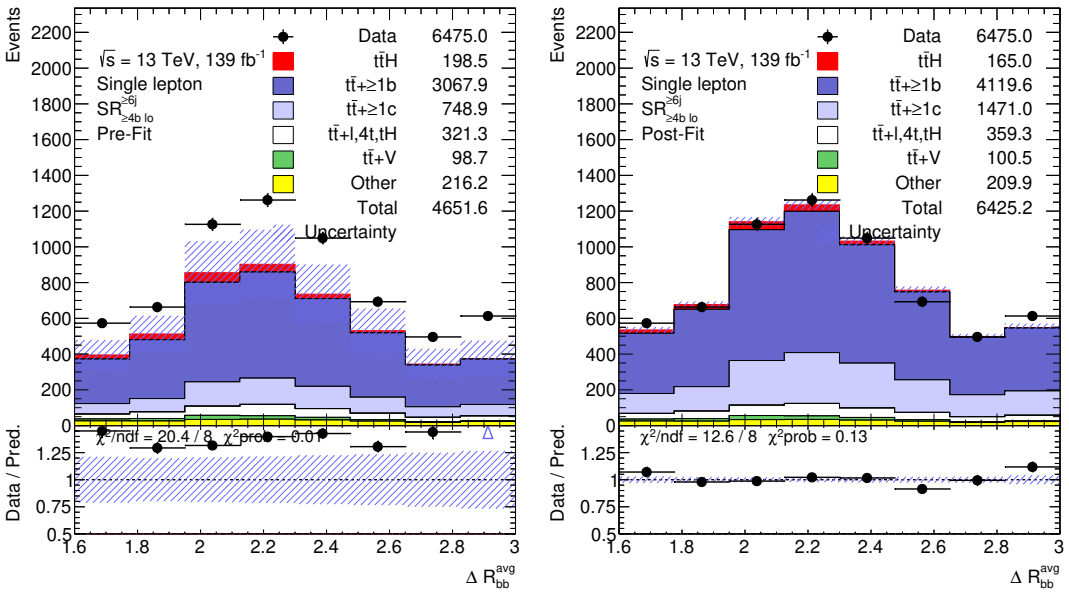


(a)

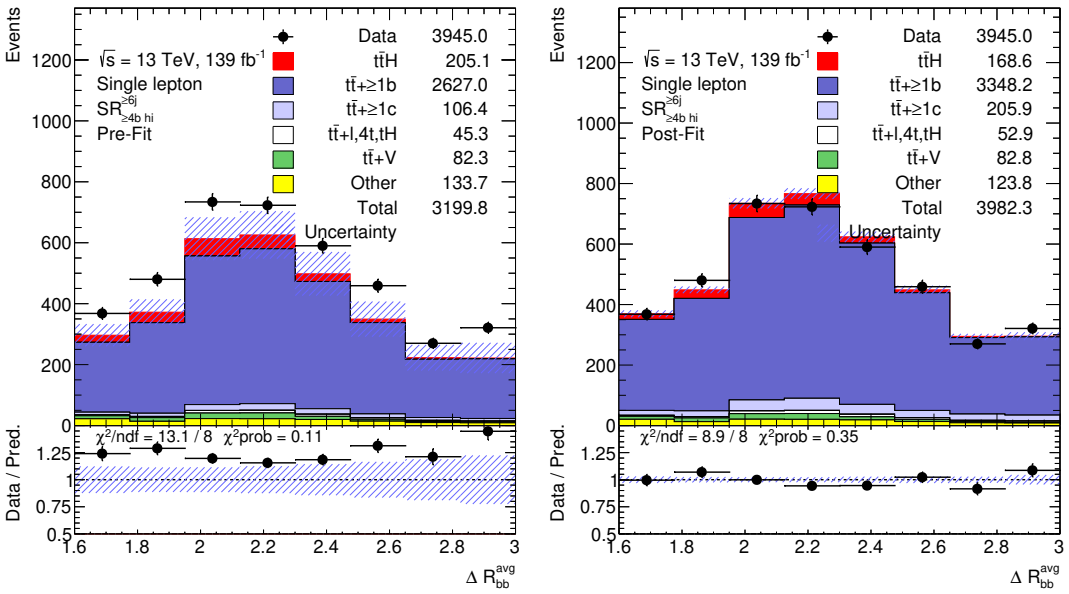


(b)

Figure D.2: Distributions of $\Delta\eta_{jj}^{max}$ pre-fit (left) and post-fit (right) in the (a) $CR^{5j}_{\geq 4b\ lo}$ and the (b) $CR^{5j}_{\geq 4b\ hi}$ regions.



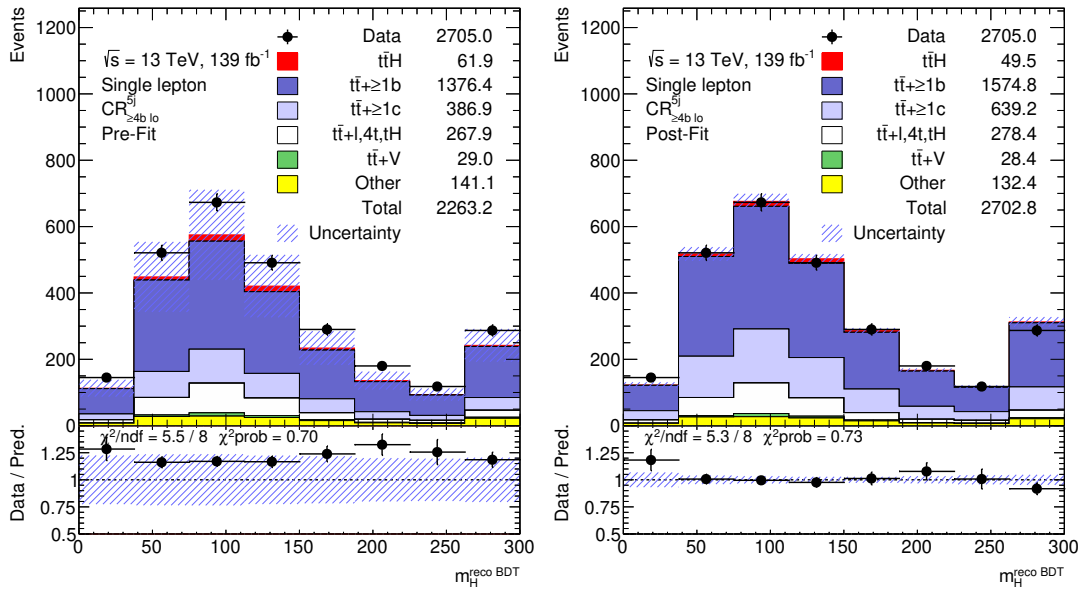
(a)



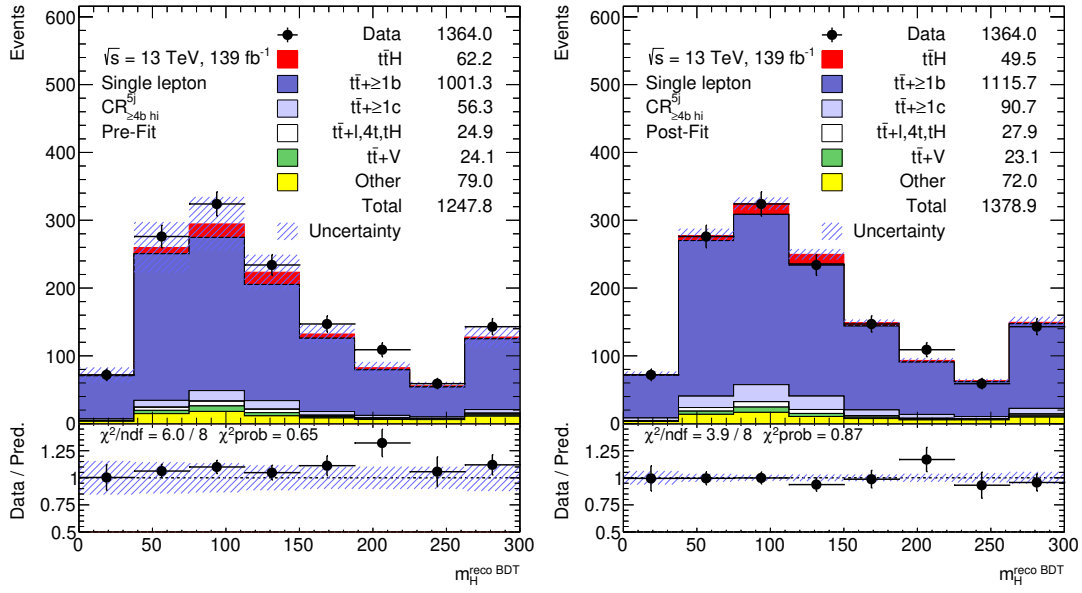
(b)

Figure D.3: Distributions of $\Delta\eta_{\mathrm{jj}}^{\mathrm{max}}$ pre-fit (left) and post-fit (right) in the (a) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{lo}}$ and the (b) $\mathrm{SR}^{\geq 6j}_{\geq 4b\ \mathrm{hi}}$ regions.

Distributions of the $m_H^{\text{reco BDT}}$ variable

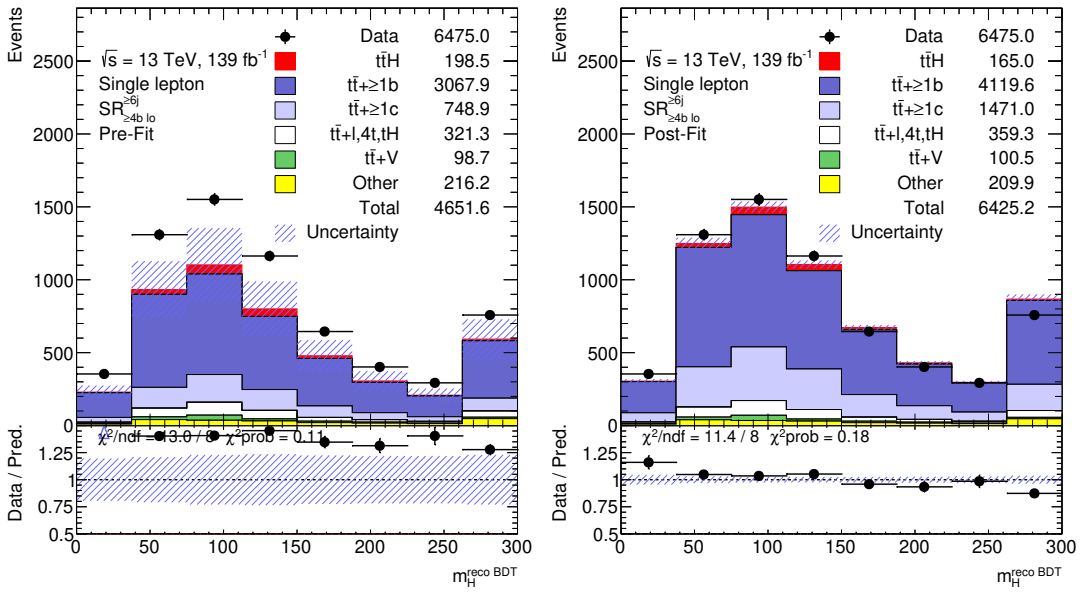


(a)

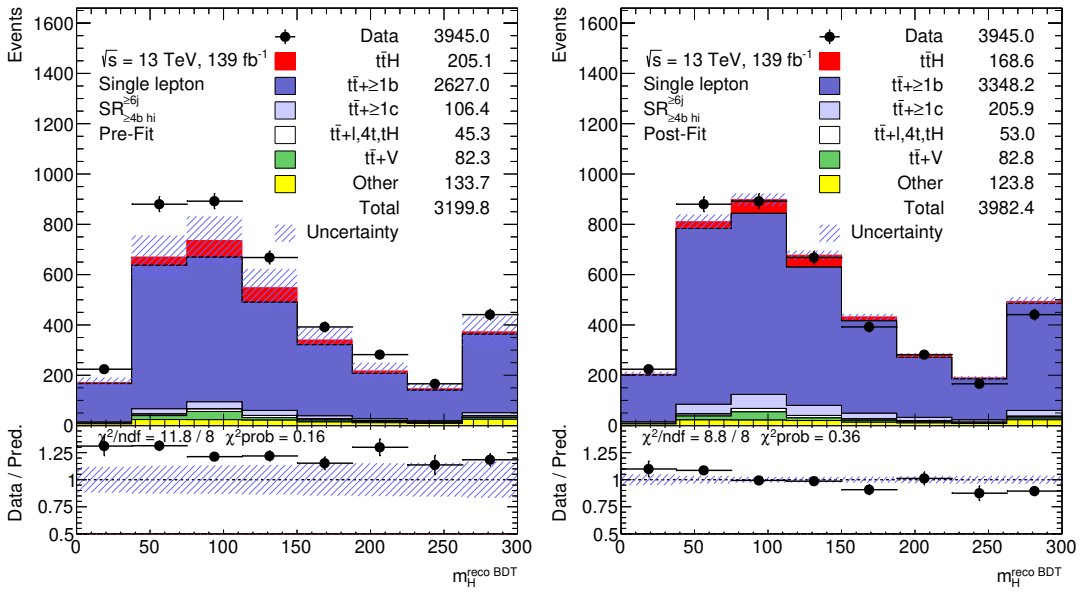


(b)

Figure D.4: Distributions of $m_H^{\text{reco BDT}}$ pre-fit (left) and post-fit (right) in the (a) $\text{CR}^{5j}_{\geq 4b\ \text{lo}}$ and the (b) $\text{CR}^{5j}_{\geq 4b\ \text{hi}}$ regions.



(a)



(b)

Figure D.5: Distributions of $m_H^{\text{reco BDT}}$ pre-fit (left) and post-fit (right) in the (a) $\text{SR}^{\geq 6j}_{\geq 4b\ \text{lo}}$ and the (b) $\text{SR}^{\geq 6j}_{\geq 4b\ \text{hi}}$ regions.